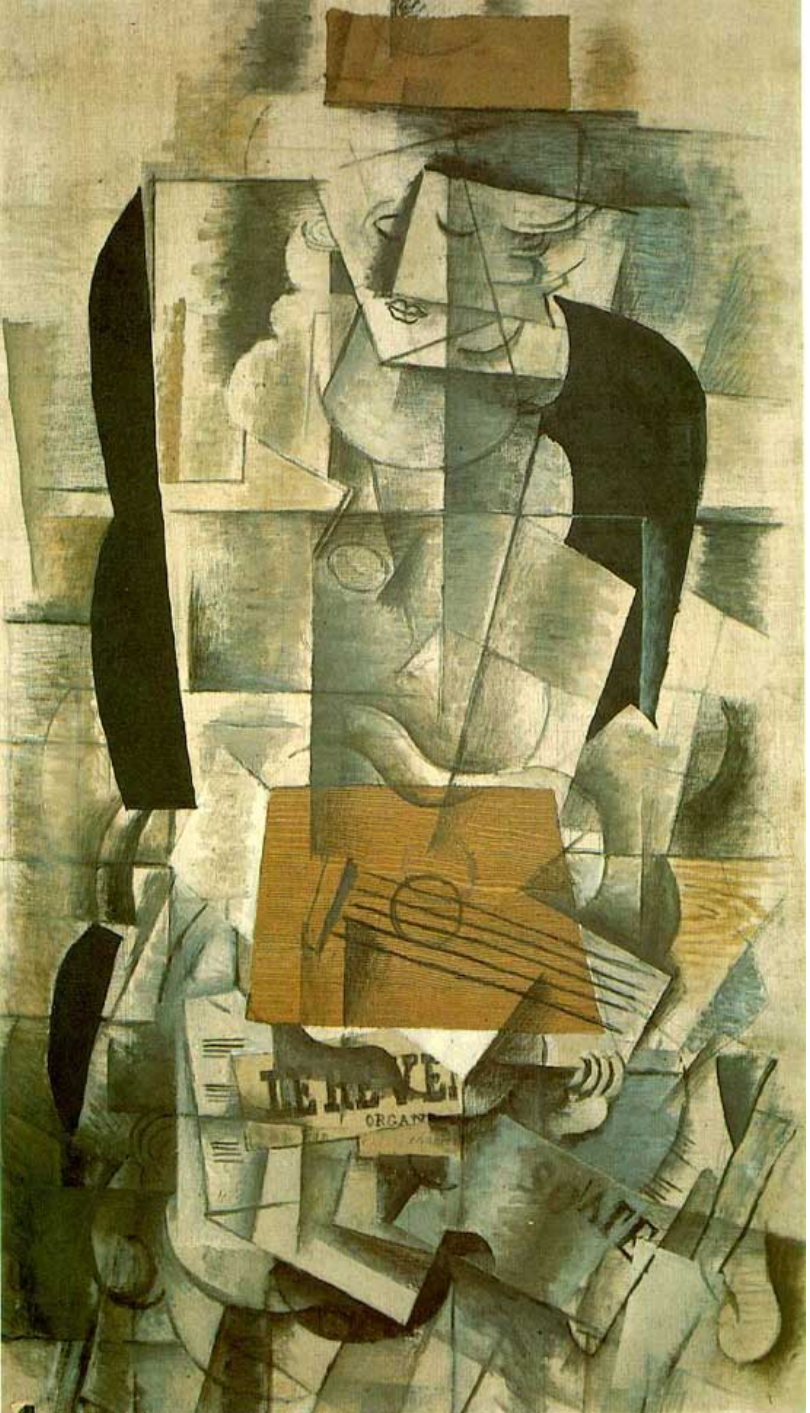
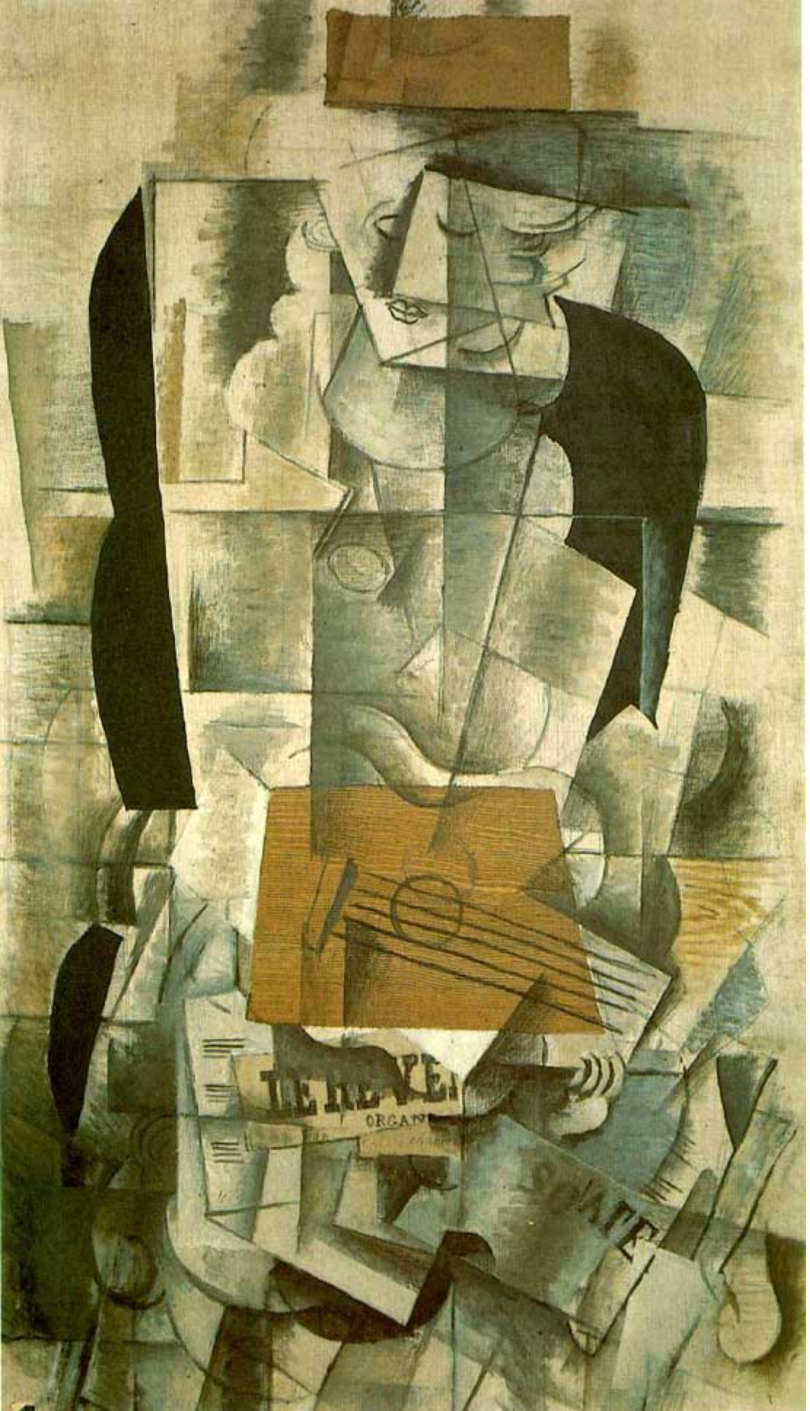
Machine Learning Models for Action Recognition

Ceyhun Burak Akgül
Vistek ISRA Vision & Boğaziçi University / Istanbul
www.vistek-isravision.com
www.busim.ee.boun.edu.tr
www.cba-research.com



Georges Braque 1913



The “WHAT”

Action Recognition

Given one or more images with one or more persons performing an action, we want to design a system recognizing the performed action.

The “HOW” – *at least part of it*

Machine Learning Model

- ❖ Learning by example
- ❖ Any statistical approach, which involves training with un/labeled data

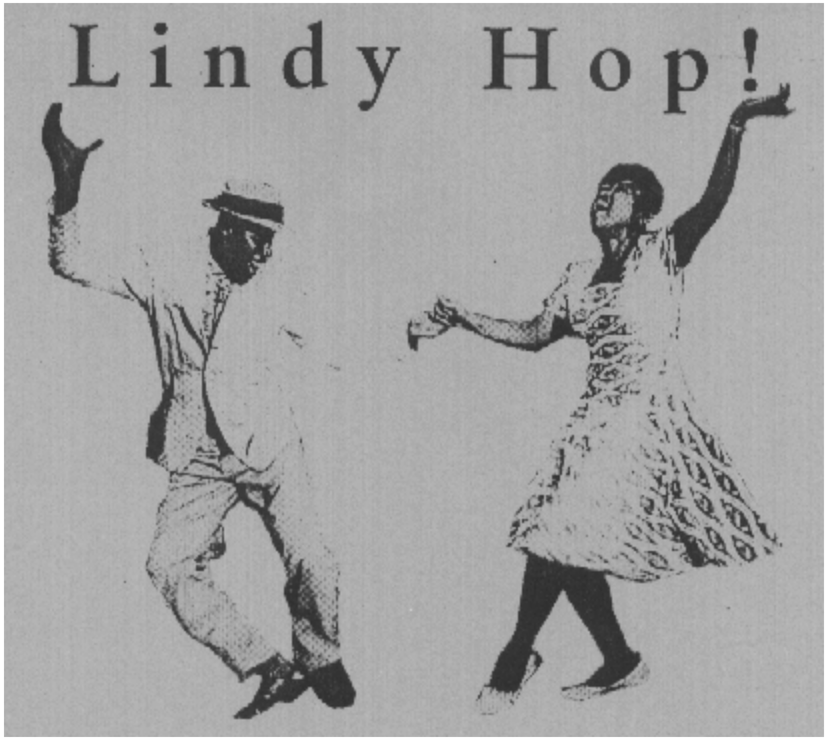
Woman with a Guitar
Georges Braque 1913

What's in an Action?

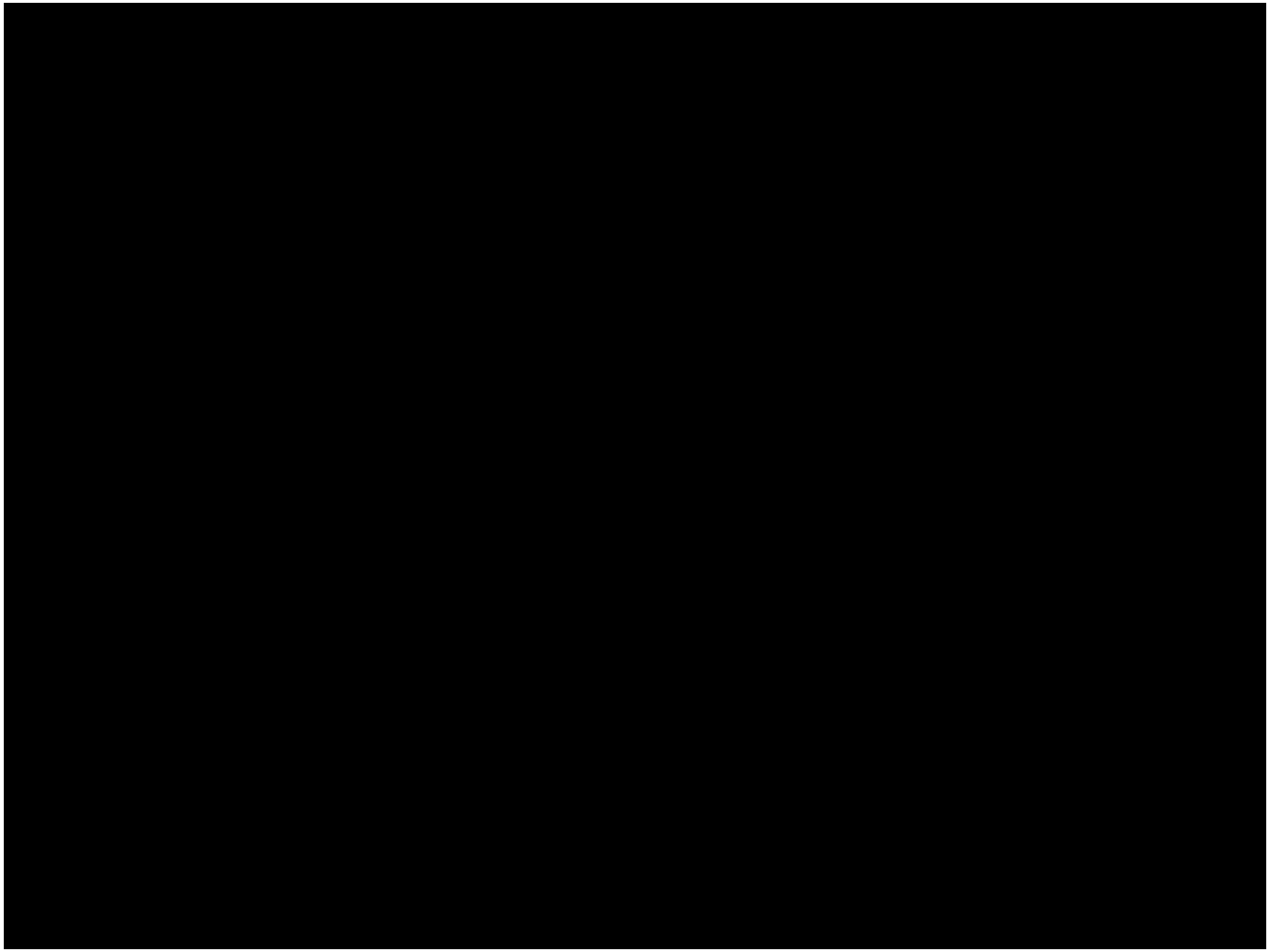
The Lindy Hop is an American dance that evolved in Harlem, New York City in the 1920s and 1930s and originally evolved with the jazz music of that time.

The Lindy Hop combines elements of both partnered and solo dancing by using the movements and improvisation of black dances along with the formal eight-count structure of European partner dances.

This is most clearly illustrated in the Lindy's basic step, the swingout.



What's in an Action?



What's in an Action?

Atomic movement
described at the limb level

Action Primitive

Basic moves:
Basic step, rock step,
triple step, hold hand,
turn, …

↓

Multiple primitives
in sequence or in combination,
possibly cyclic whole-body movement

Action

Solo Swingout:
Rock step -> step -> triple step &
turn -> step -> step -> triple step

↓

Multiple actions
in sequence or in combination,
possibly involving multiple
persons and/or objects

Activity

Swing out with a partner

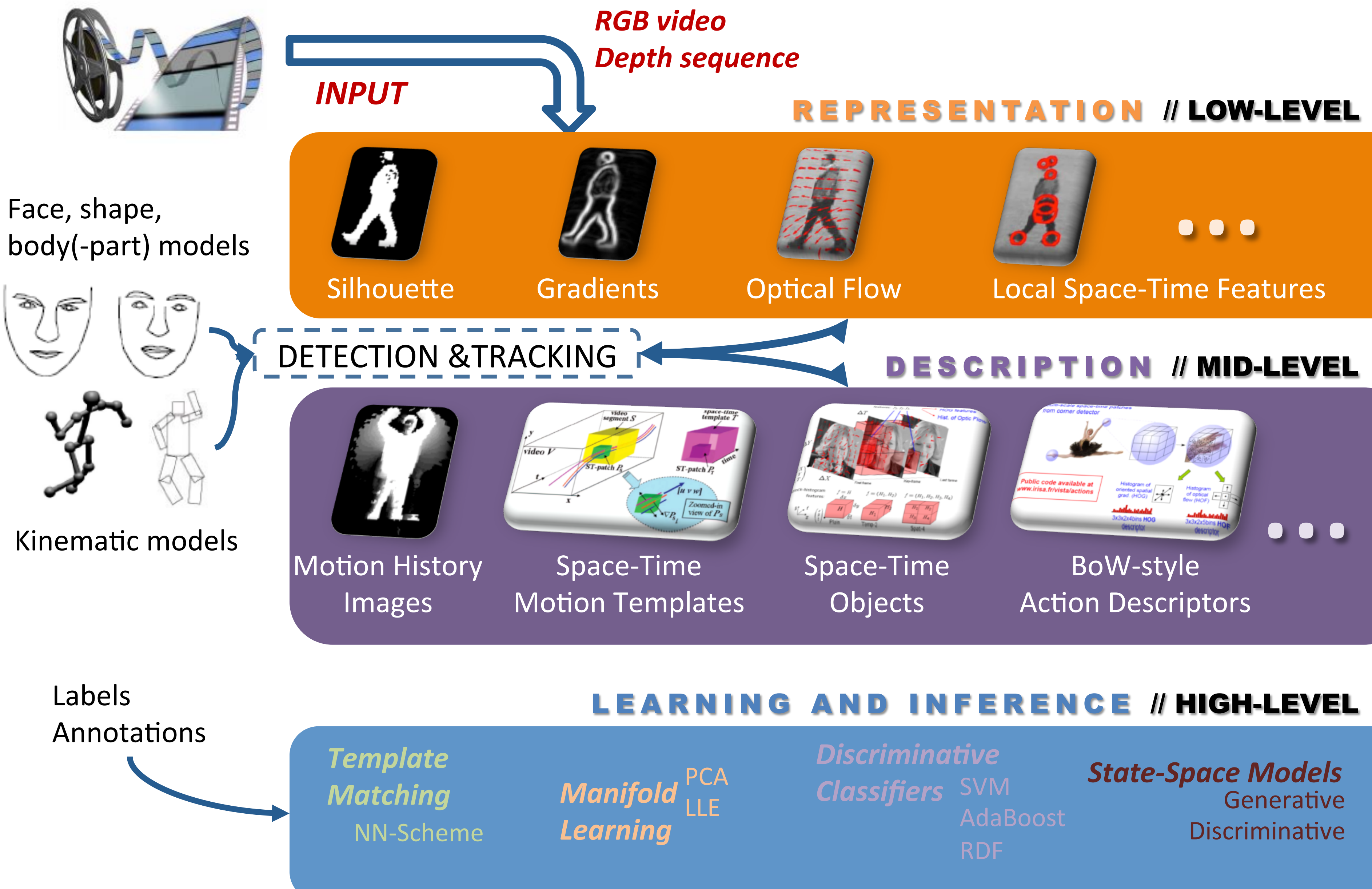
↓

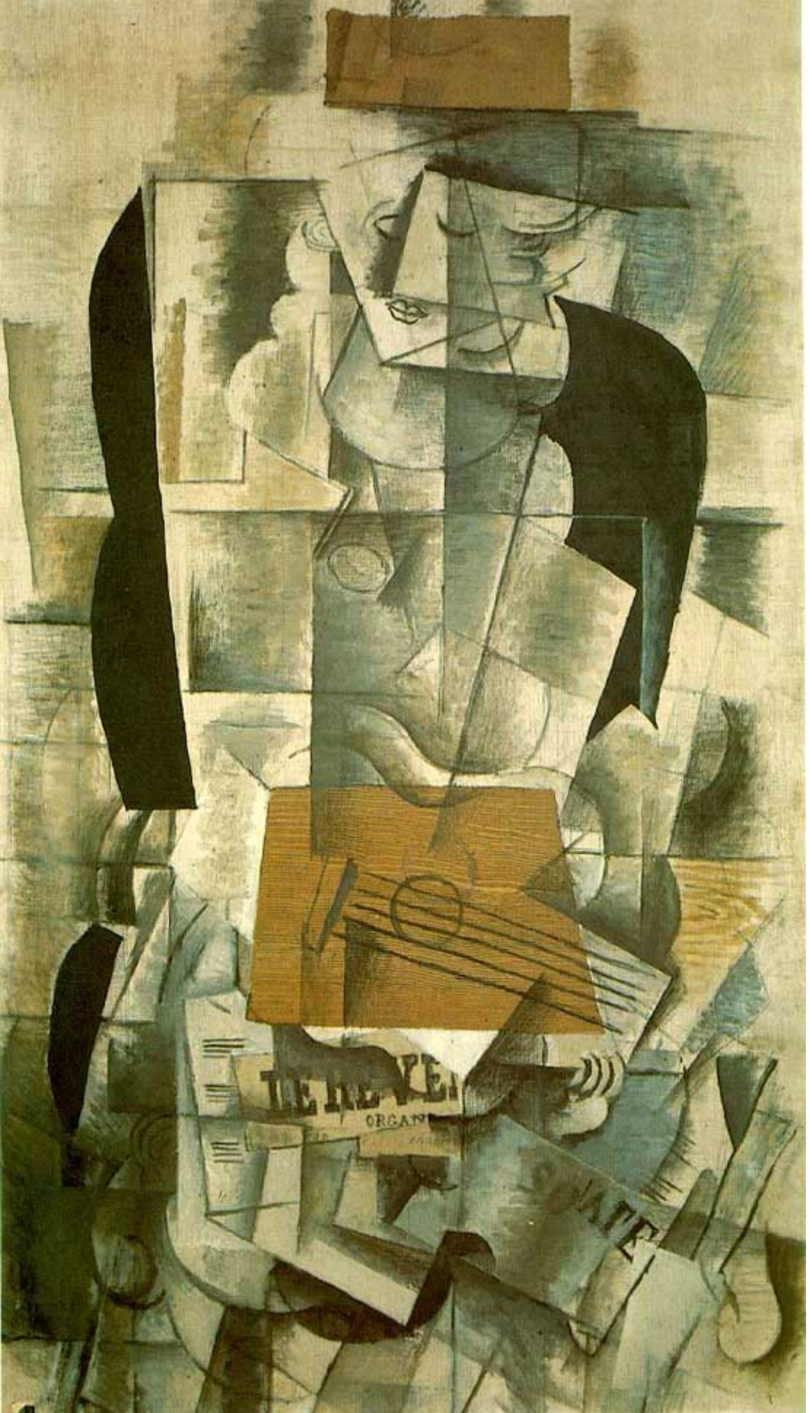
Series of multiple activity
instancesinvolving different
people and objects in a specific context

Event

The Lindy Hop Party!

The Blueprint Action Recognizer…





Outline(*)

Challenges

Surveys

Datasets

A Parade of ML Models

- *The Nearest Neighbor Scheme*
- *Manifold Learning*
- *Discriminative Classifiers*
- *State-Space Models*

Variations on the Theme

- *Mining Action Data*
- *Use of Context*

Concluding Remarks

(*) The full set of slides can be downloaded from
http://www.cba-research.com/pdfs/MLM4AR_DemAAL2013_CBAkgul.pdf

Woman with a Guitar
Georges Braque 1913

why is it difficult?

CHALLENGES

Class Definitions and Variability
Environment and Recording Settings
Spatio-Temporal Variability
Real-Time Recognition
On-the-Fly Recognition
Training Data Collection and Labeling
Evaluation and Benchmarking

Challenges – 1/4

Class Definitions and Variability

Basic
Walking
Jogging
Running
Boxing
Hand waving
Hand clapping
…

Daily Living
Getting out of bed
Watching TV
Reading a book
Using computer
Eating meal
Drinking
…

Outdoor
Walking alone
Meeting w/ others
Window shopping
Fighting
Leaving luggage behind
…

Action

increasingly more complex and variable…
anthropometric differences
multiple objects and people
contextual differences

Activity

Challenges – 2/4

Environment and Recording Settings

Issues	Consequences
• Static vs. Dynamic backgrounds • Occlusions • Lighting conditions • Recording rate and resolution • Recording modality	• Person detection and tracking • Action detection and segmentation • Level of detail for understanding • Choice of method

Challenges – 3/4

Spatio-Temporal Variability

Issues	Consequences
• Pose differences • Moving camera • Execution time and rate	• View invariance required • Person detection and tracking • Action detection and segmentation • Temporal effects: remove or take into account?

Challenges – 4/4

Other Challenges

- Real-Time Recognition
- On-the-Fly Recognition
- Training Data Collection and Labeling
  - Reliable and objective annotations required for learning
  - Large and varied training and test data for all classes required for generalization
- Evaluation and Benchmarking
  - Common realistic benchmarks required to compare methods

who has done what?

SURVEYS

Taxonomies, taxonomies …

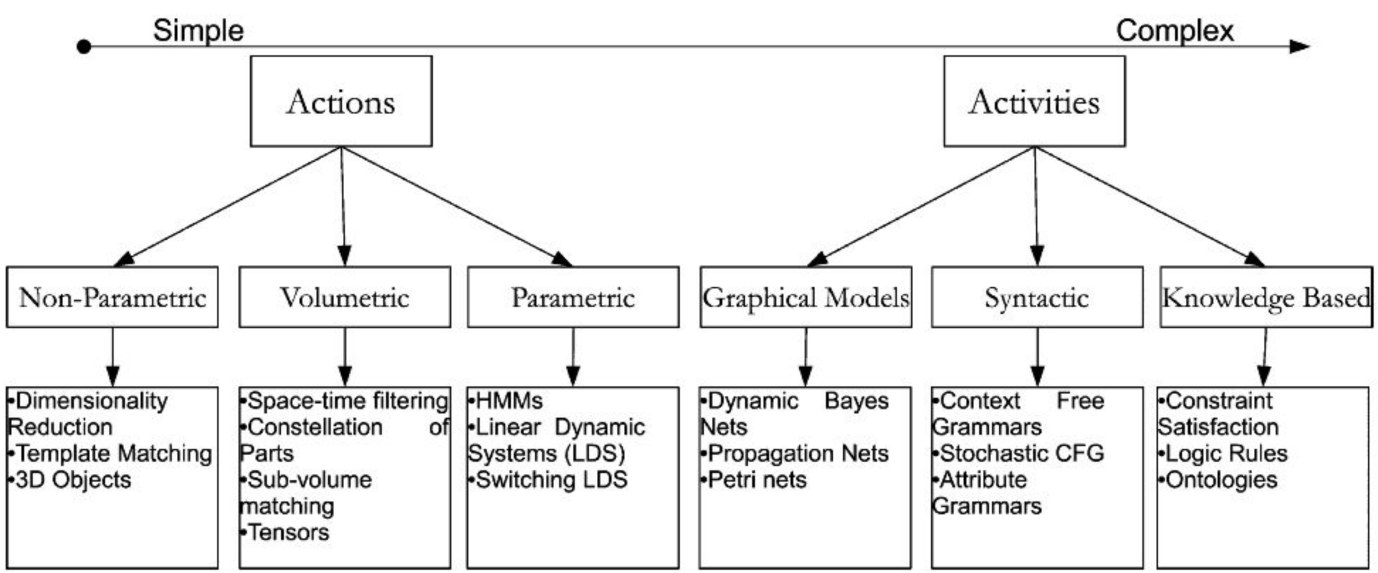
[Moeslund et al., 2006]

- 352 papers covered for the period 2000-2006
- Functional taxonomy: Initialization, tracking, pose estimation, tracking

[Turaga et al., 2008]

- 144 papers covered

Turaga et al.'s methodological taxonomy



Taxonomies, taxonomies …

[Poppe, 2010]

- 180 papers covered
- Representation and classification aspects treated separately

[Weinland et al., 2011]

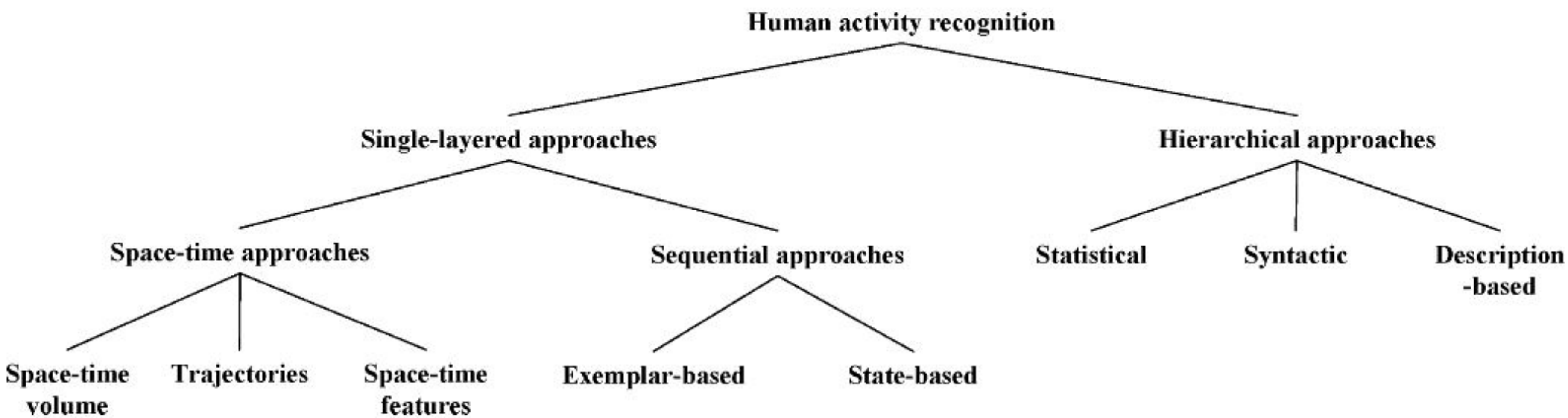
- 153 papers covered
- Focused on representational aspects (spatial vs. temporal) as well as action segmentation and view invariance
- Classification and Learning aspects not discussed

Taxonomies, taxonomies …

[Aggarwal and Ryoo, 2011]

- 102 papers covered

Aggarwal and Ryoo’s hierarchical approach-based taxonomy



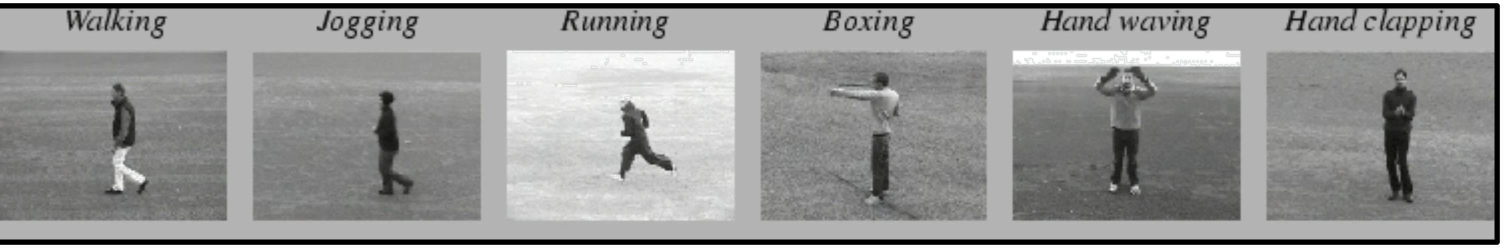
where to train and test?

DATASETS

- **The Usual Suspects**
- **Surveillance Datasets**
- **The Wild Ones**
- **Datasets for ADL**
- **Rising Stars: RGBD Datasets**

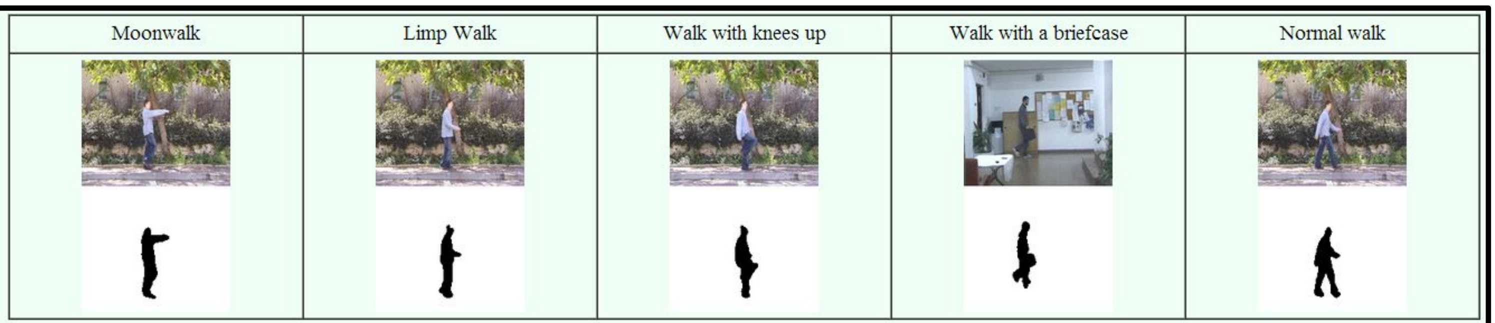
The Usual Suspects

KTH



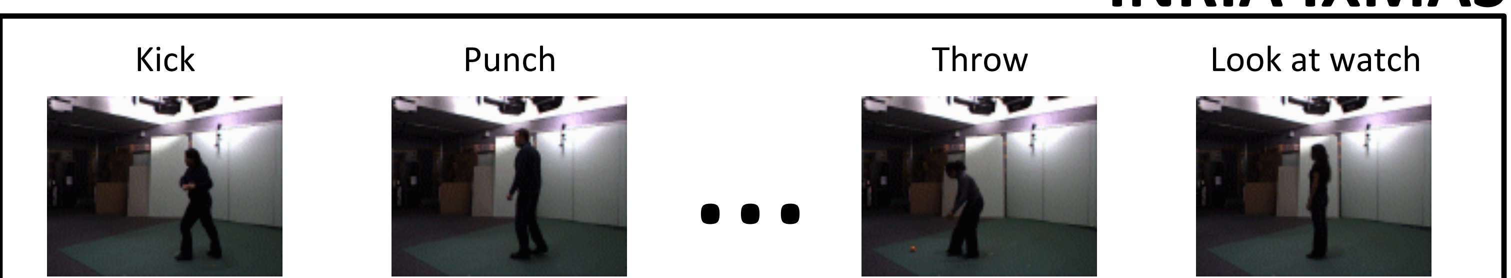
6 actions
25 subjects
Simple background

Weizmann



10 actions
Class variations
Varied background

INRIA IXMAS



11 actions
12 subjects
Controlled env.
Gaming scenario

Surveillance Datasets

PETS

- **Performance Evaluation of Tracking and Surveillance Challenge** (since 2000)
- Focused on crowd surveillance characteristics/events within a real-world environment
- Person count and density estimation – People Tracking – Flow Analysis and Event Recognition

CAVIAR

- **CAVIAR** project video clips collected at public spaces (entrance lobby and shopping mall) using a wide angle lens
- Activities: people walking alone, meeting with others, window shopping, entering and exiting shops, fighting and passing out and leaving a package in a public place.

SDHA

- **Semantic Description of Human Activities:** Three Challenges in ICPR 2010
- **Interaction Challenge:** High-level interactions between two humans, e.g., hand-shake and push
- **Aerial View Challenge:** Simple one-person actions taken from a low-resolution far-away camera
- **Wide Area Challenge:** Monitor human activities with multiple cameras observing a wide area

ViSOR

- **Vi**deo **S**urveillance **O**nline **R**epository
- Diverse environments and settings: outdoor, indoor,
- Object-level and action/activity-level meta-data available

The Wild Ones – 1/4

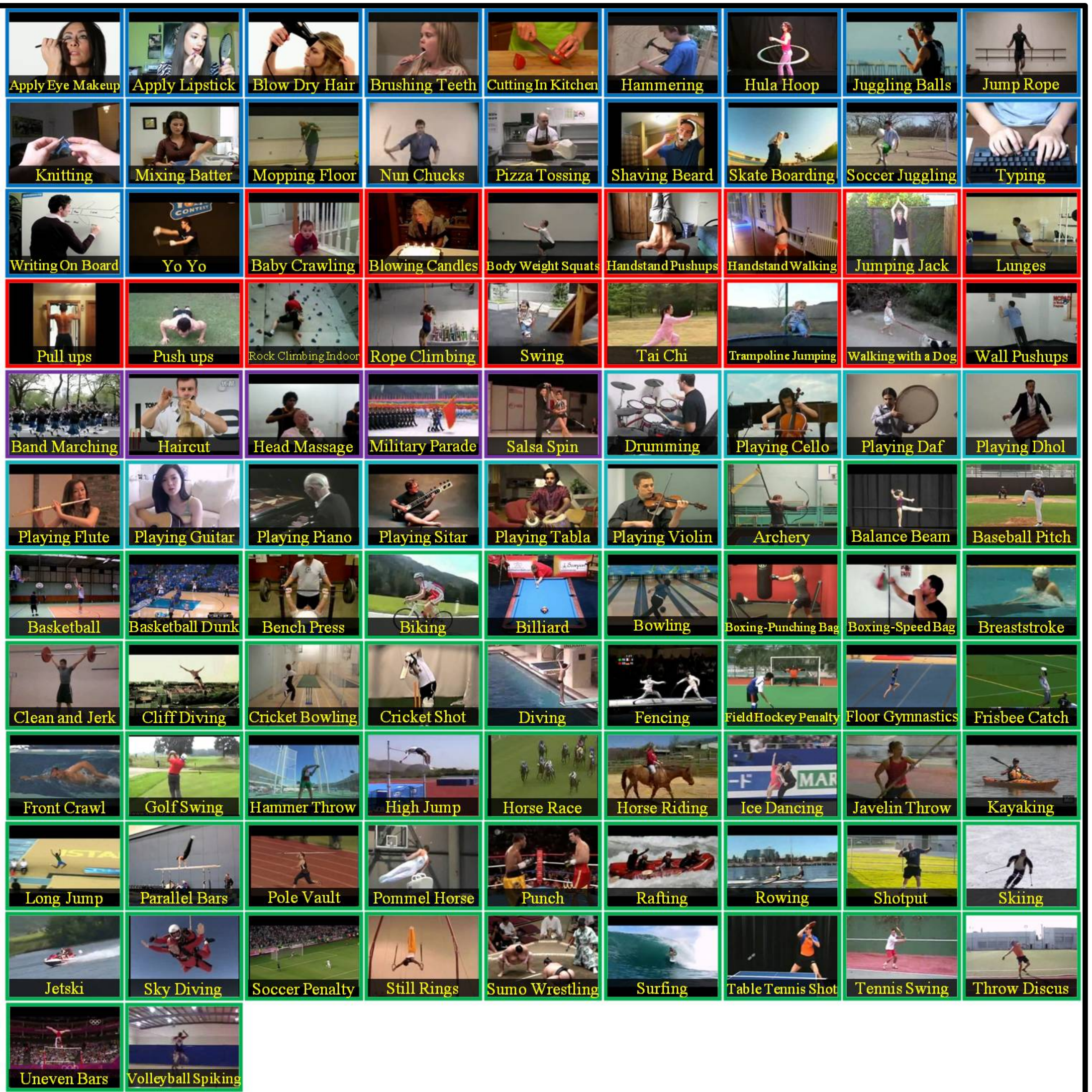
Hollywood2

12 classes of human actions and 10 classes of scenes
3669 video clips from 69 movies
Approximately 20.1 hours of video
Comprehensive benchmark in realistic and challenging settings

Hollywood2
UCF101
HMDB
ActionBank

The Wild Ones – 2/4

UCF101



101 action categories:
(extension of UCF50)
(1) Human-Object Interaction
(2) Body-Motion Only
(3) Human-Human Interaction
(4) Playing Musical Instruments
(5) Sports.

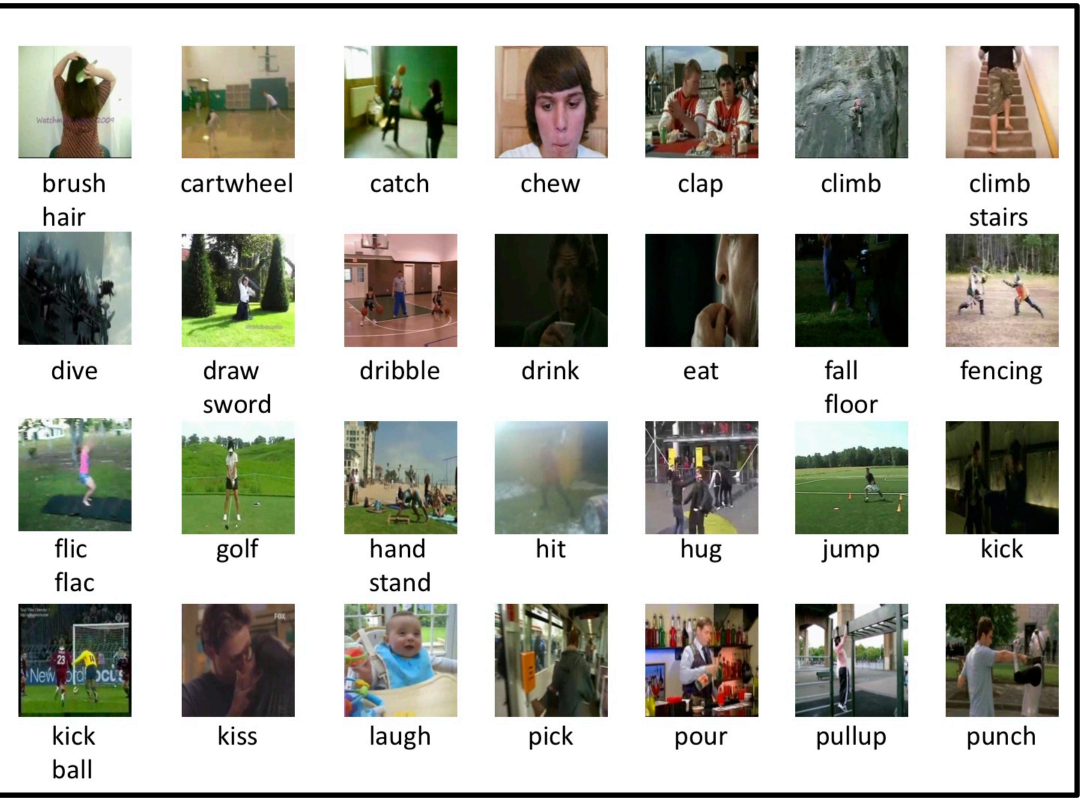
13320 videos from YouTube

Large diversity:
actions classes, large variations in camera motion, object appearance and pose, object scale, viewpoint, cluttered background, illumination conditions, etc.

No actors

The Wild Ones – 3/4

HMDB



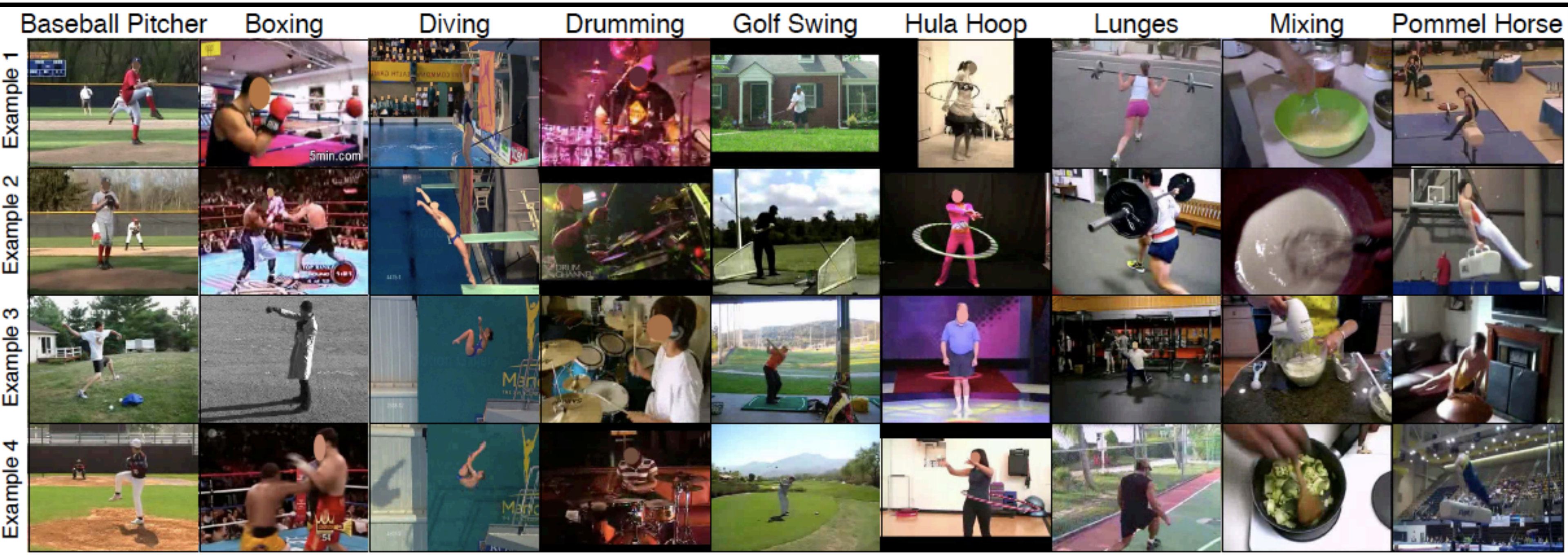
51 action categories:
(1) General facial actions
(2) Facial actions with object manipulation
(3) General body movements
(4) Body move'ts with object interaction
(5) Body move'ts for human interaction

6849 clips from the Prelinger archive, YouTube and Google videos
(minimum 101 clips per category)

The Wild Ones – 4/4

Hollywood2
UCF101
HMDB
ActionBank

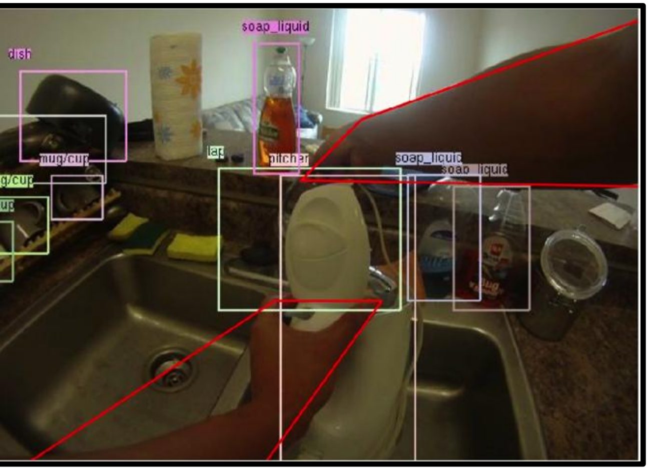
ActionBank



A combination of KTH, UCF Sports, UCF50, HMDB51

Datasets for ADL [Activities of Daily Living] – 1/2

ADLs differ from typical actions in that they can involve long-scale temporal structure (making tea can take a few minutes) and complex object interactions (a fridge looks different when its door is open)



UCI-ADL

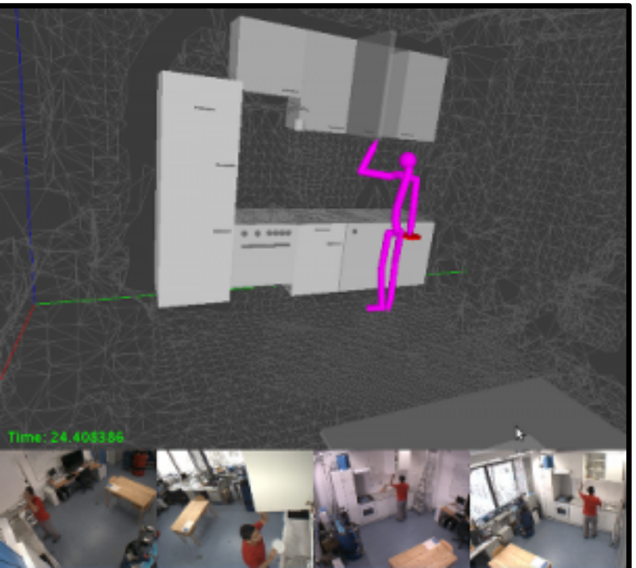
1 million frames of dozens of people performing ADL

Annotated with activities, object tracks, hand positions, and interaction events.

YouCook

88 YouTube cooking videos (various recipes) from third-person viewpoint

Frame-by-frame object and action labels



TUM-Kitchen

Observations of several subjects setting a table in different ways.

Video data
Motion capture data
RFID tag readings
Magnetic sensor data
Detailed action labels

Datasets for ADL [Activities of Daily Living] – 2/2

ADLs differ from typical actions in that they can involve long-scale temporal structure (making tea can take a few minutes) and complex object interactions (a fridge looks different when its door is open)

UESTC Senior Home Monitoring Dataset

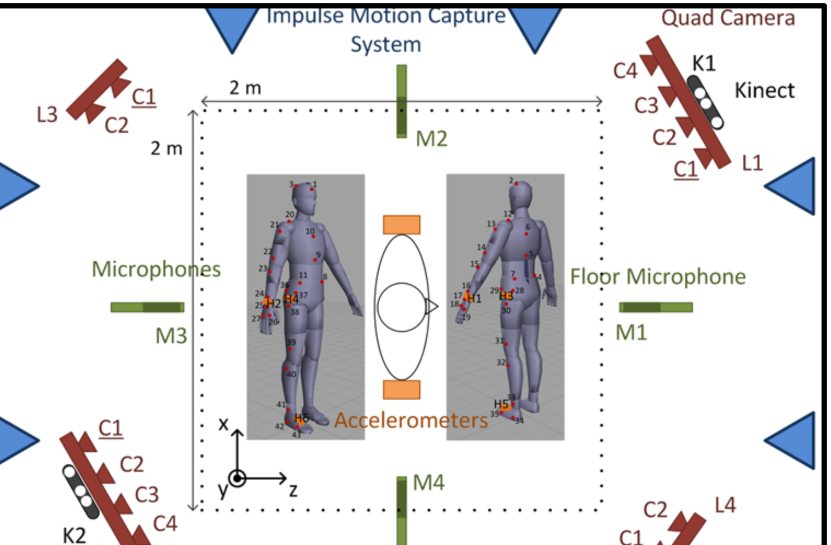
12 types of human actions:
drinking, eating meals, eating snacks, getting out of bed, going to bed, sleeping, smoking, walking, playing mahjong, washing face, washing feet and watching TV

Performed by 6 seniors in their own rooms
4 month long data collection
10 days per recording for each senior
Approximately 1.8TB data
(25fps, 360x288 pixels, Xvid MPEG-4 Codec)

Rising Stars: RGBD Datasets – 1/5

Berkeley MHAD
MSR
Cornell
LIRIS
WorkoutSU-10

Berkeley MHAD



11 actions by 7 male and 5 female subjects
(23-30 years except one elderly)
5 repetitions per subject per action
660 action sequences, 82 minutes total recording time

(1) Movements in both upper and lower extremities
(2) Actions with high dynamics in upper extremities
(3) Actions with high dynamics in lower extremities

Simultaneously captured by five different systems: optical motion capture system, four multi-view stereo vision camera arrays, two Microsoft Kinect cameras, six wireless accelerometers and four microphones.

Rising Stars: RGBD Datasets – 2/5

Microsoft Research (MSR) Datasets

MSRGesture3D

Depth sequences captured by Kinect

12 dynamic American Sign Language (ASL) gestures

10 people, 2-3 times per subject per gesture class, 336 depth sequences

MSRDailyActivity3D

Depth, RGB, and skeletal data sequences captured by Kinect (RGB and depth not synchronized)

16 activities: drink, eat, read book, call cellphone, write on a paper, use laptop, …

10 subjects, 2 times per subject per activity (one in standing, the other in sitting position)

MSRAction3D

Depth and skeletal joint data sequences captured by Kinect-like device

20 general action classes

10 subjects, 2-3 times per subject per activity 567 depth sequences

MSRC-12

Depth sequences and skeletal data captured by Kinect

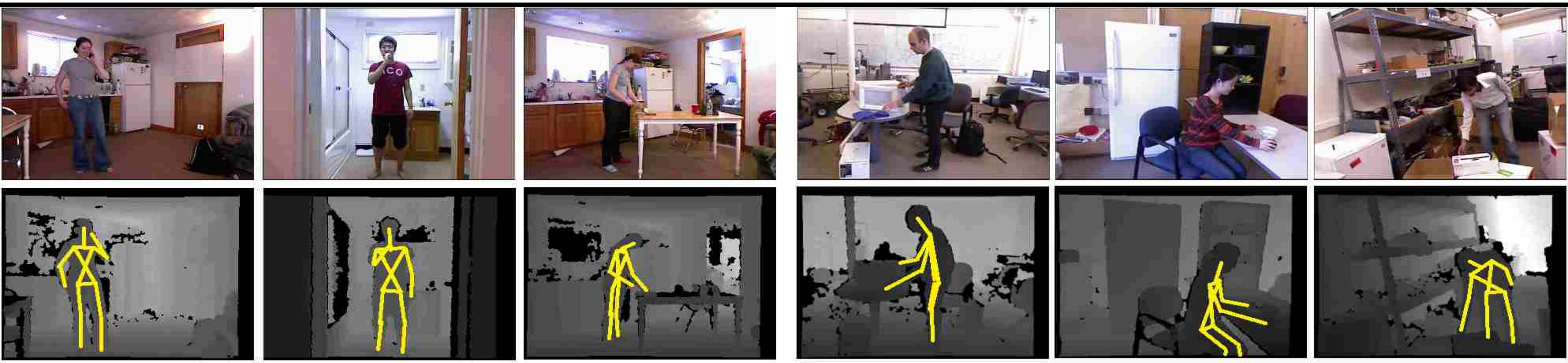
12 gesture classes from a 1st person shooter video game

30 people, 6244 gesture instances in 594 sequences (6hrs 40min)

Rising Stars: RGBD Datasets – 3/5

Berkeley MHAD
MSR
Cornell
LIRIS
WorkoutSU-10

Cornell Activity Datasets



CAD-60

60 RGB-D videos and tracked skeletons

4 subjects: 2 male, 2 female (one left-handed)

5 different environments: office, kitchen, bedroom, bathroom, and living room

12 activities: rinsing mouth, brushing teeth, wearing contact lens, talking on the phone, drinking water, …

CAD-120

120 RGB-D videos of long daily activities

4 subjects: 2 male, 2 female (one left-handed)

10 high-level activities: making cereal, taking medicine, un/stacking objects, microwaving …

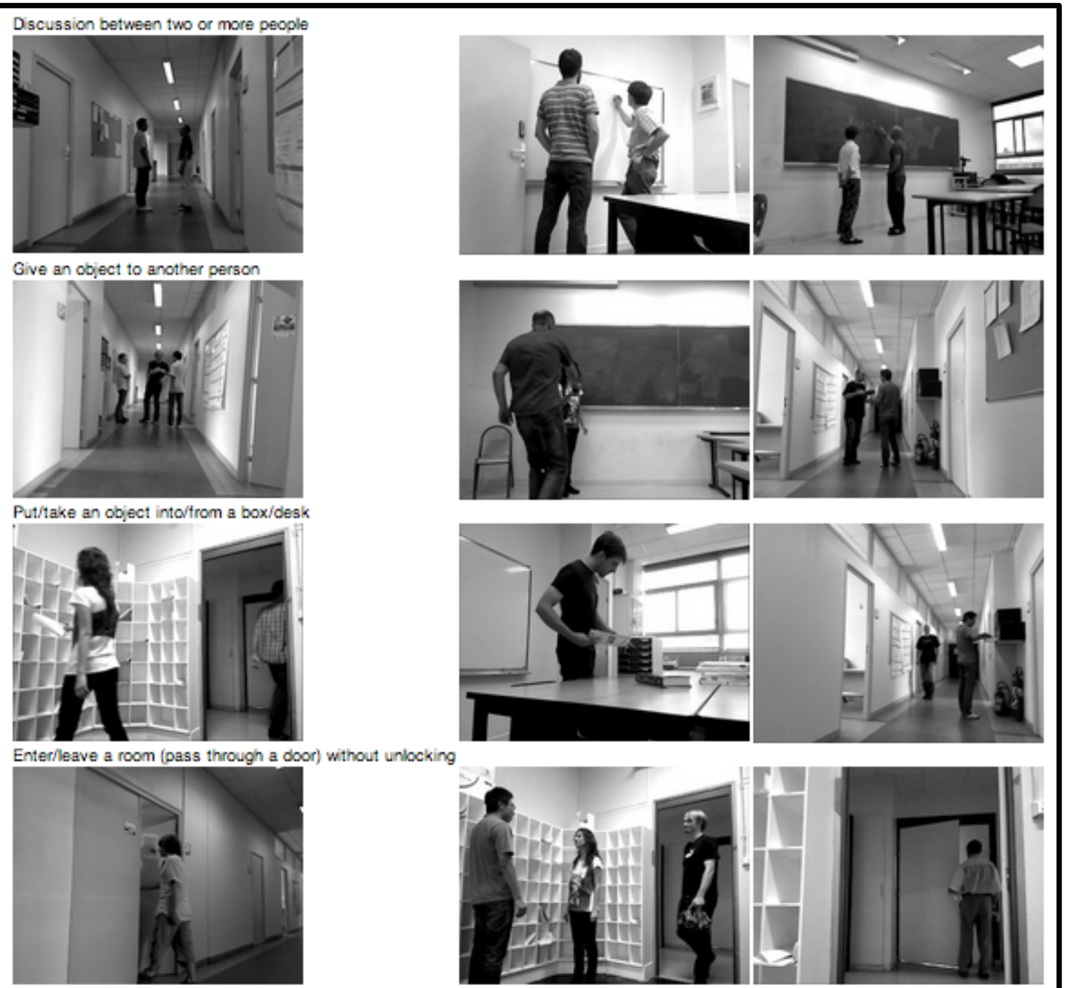
10 sub-activity (action) labels: reaching, moving, pouring, eating, drinking, …

12 object affordance labels: reachable, movable, pourable, containable, …

Rising Stars: RGBD Datasets – 4/5

Berkeley MHAD
MSR
Cornell
LIRIS
WorkoutSU-10

LIRIS Human Activities Dataset



RGB, grayscale and depth sequences

RGB-D videos of various ADL:
discussing, phone calls, giving an item, ...

Fully annotated with spatial and temporal positions in video

Originally shot for the ICPR-HARL 2012 competition

Rising Stars: RGBD Datasets – 5/5

WorkoutSU-10

Gesture out-come/Code	Descriptive Instruction	Image
SL Balance with Hip Flexion	• Flex your hip of your non-weight bearing leg up to 90 degrees, bend your knee, and hold. • Use your core & lower extremity muscles to control your center of mass to maintain your balance.	
SL Balance-Trunk Rotation	• Raise your arms to chest height and clasp your hands together. • Slowly rotate your trunk to one side a comfort-able distance, return to the starting position, and then rotate your trunk in the other direc-tion. • Use your core & lower extremity muscles to control your center of mass to maintain your balance.	
Lateral Stepping	• Slightly bend your knees and begin stepping to the side keeping your toes facing straight ahead. • Use your core & lower extremity muscles to control your center of mass to maintain your balance. • Perform this for a specific number of steps then return back in the other direction.	
Thoracic Rotation – Bar on shoulder	• Assume standing position with bar across shoulders. • Rotate your trunk to one side. • Hold 30 (s) at end range; then slowly release stretch.	
Hip Adductor Stretch	• Shift your weight over one leg by bending your knee and straighten the opposing leg to be stretched. • You should feel a stretch on the inside as-pect of your thigh and groin of the straight leg. • Hold 30 (s) at end range; then slowly re-lease the stretch.	

Depth sequences and skeletal data captured by Kinect

Balance Exercises
Stretching and Flexibility Exercises
Strengthening Exercises

10 therapeutic action classes in 3 broad categories

15 participants
10 repetitions per subject per class, 1200 instances in total

Recorded in the context of the ViPSafe Project on elderly monitoring (Sabancı University and Vistek ISRA Vision)

The Nearest Neighbor Scheme
Manifold Learning
Discriminative Classifiers
State-Space Models

the toolbox…

MACHINE LEARNING MODELS

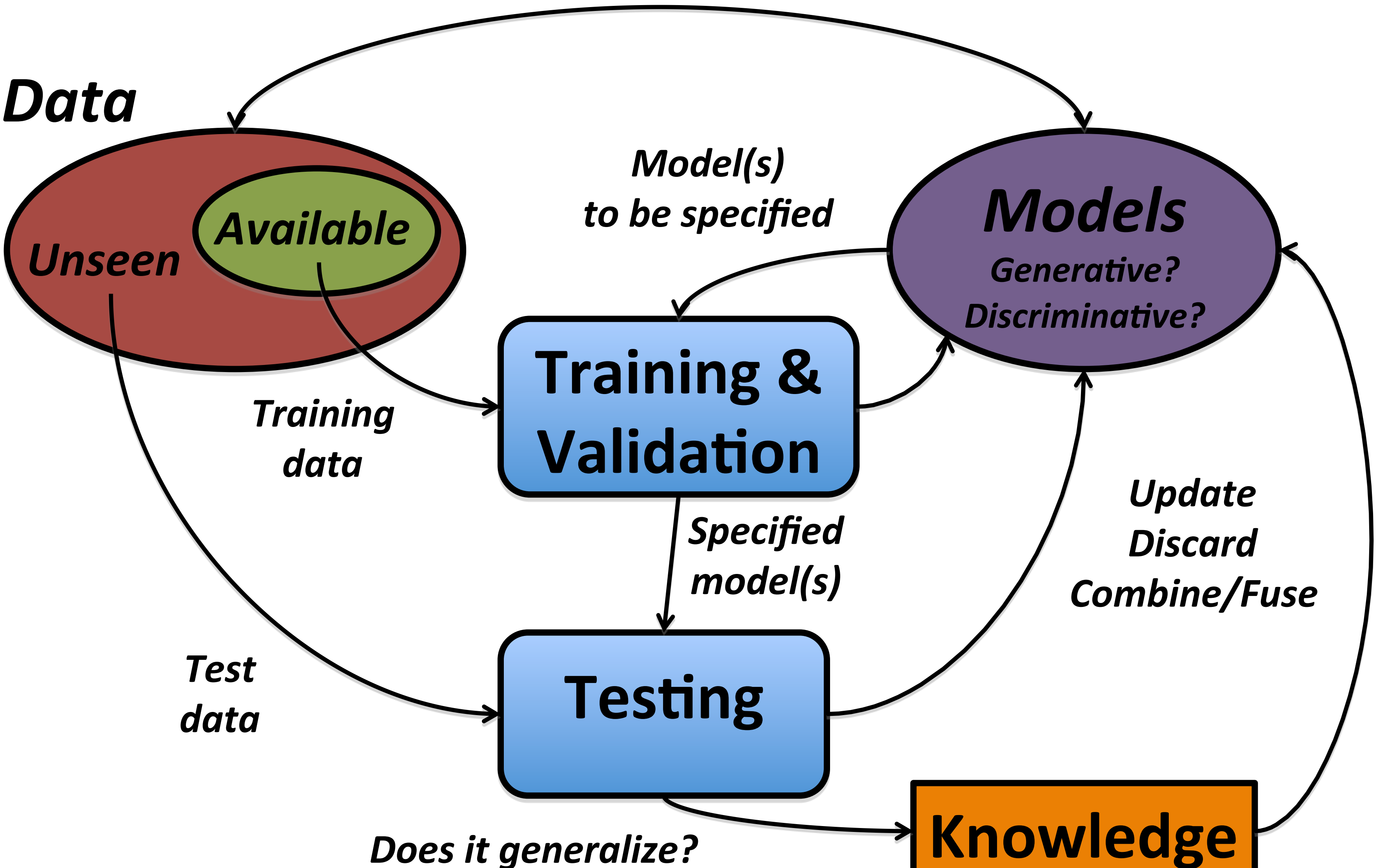
The Bayes Classifier

$$\boxed{C^* = \operatorname{argmax} P(C|D)}$$

C: action class
D: description of the observed visual data
P(C|D): posterior probability of class **C** having observed description **D**

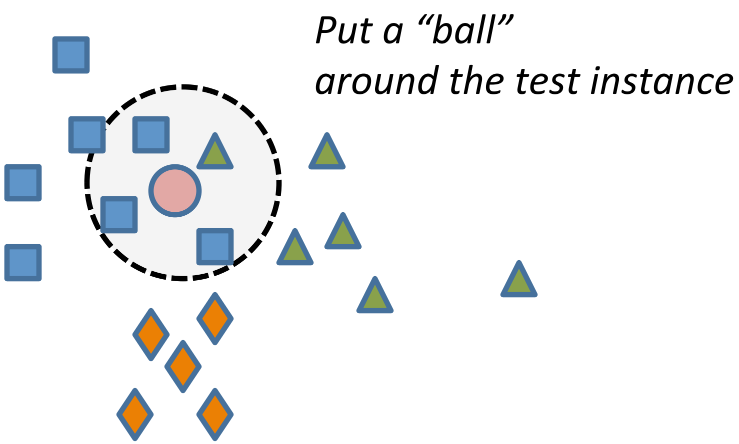
All machine learning models try to approximate this formula in one way or the other

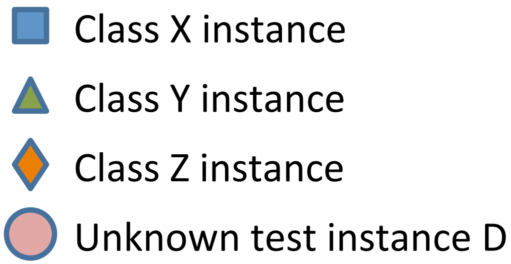
The Machine Learning Pipeline…



The Nearest Neighbor Scheme

The Nearest Neighbor Scheme

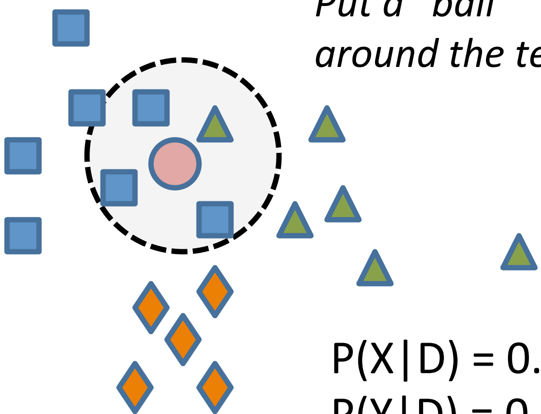




The Nearest Neighbor Scheme

- Class X instance
- Class Y instance
- Class Z instance
- Unknown test instance D

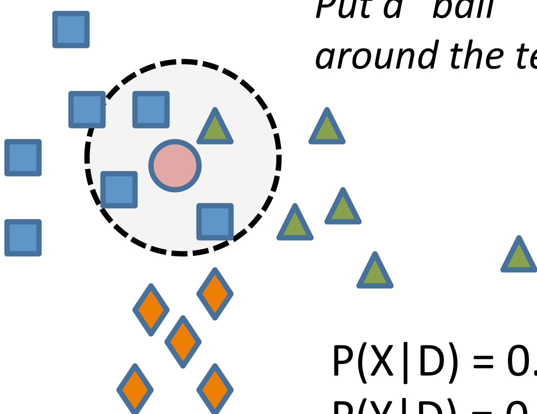
Put a "ball" around the test instance



P(X|D) = 0.75
P(Y|D) = 0.25
P(Z|D) = 0.00

Assign D to class X

The Nearest Neighbor Scheme



Put a "ball"
around the test instance

Class X instance

Class Y instance

Class Z instance

Unknown test instance D

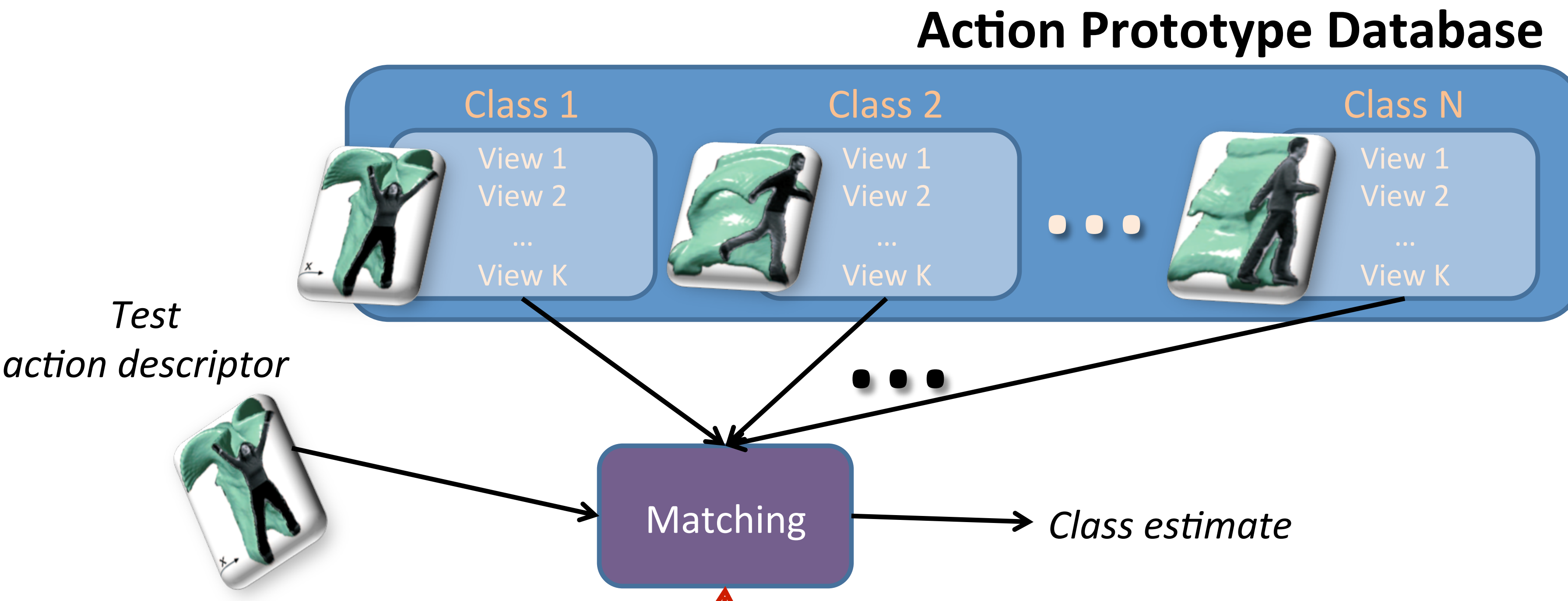
P(X|D) = 0.75
P(Y|D) = 0.25
P(Z|D) = 0.00

Assign D to class X

In which space should we put the ball?

- **Action description**
- **Matching measure**

The Nearest Neighbor Scheme



Distance Measure

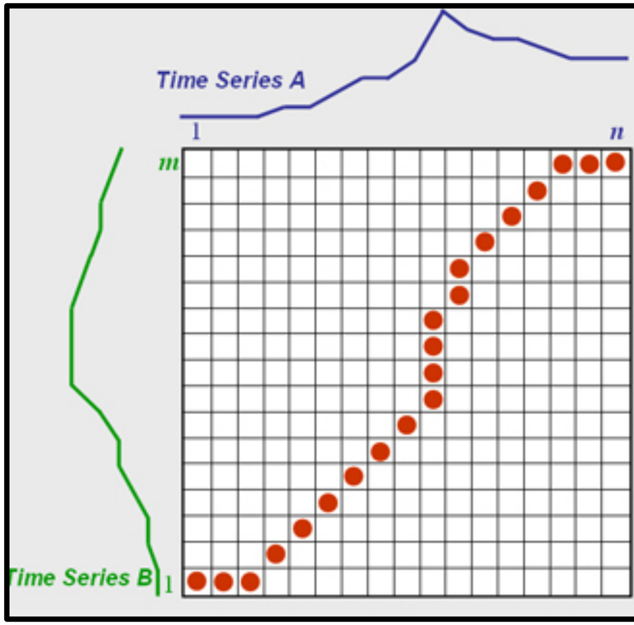
Minkowski
Deformable
Dynamic Time Warping
...

Space-time shapelets [Batra et al. 2008]
Space-time shapes [Gorelick et al. 2007]
Motion History Images [Bobick and Davis 2001]
Action MACH [Rodriguez et al. 2008]
Shape Motion Prototype Trees [Lin et al. 2009]

The Nearest Neighbor Scheme

Dynamic Time Warping (DTW)

- DTW computes a nonlinear (space-)time normalization between a template and a test vector
- Vectors could be of different length
- Better capture intrapersonal variations in gait than linear warping
- Computation based on dynamic programming
- Can be speeded-up by using certain (spatio-) temporal consistency constraints.



Non-parametric matching of shape sequences [Veeraraghavan et al. 2005]
Function space of an activity [Veeraraghavan et al. 2006]
Deformable action templates [Yao and Zhu 2009]

Manifold Learning

In which space
should we put the "ball"?

- **Action description**
- **Matching measure**

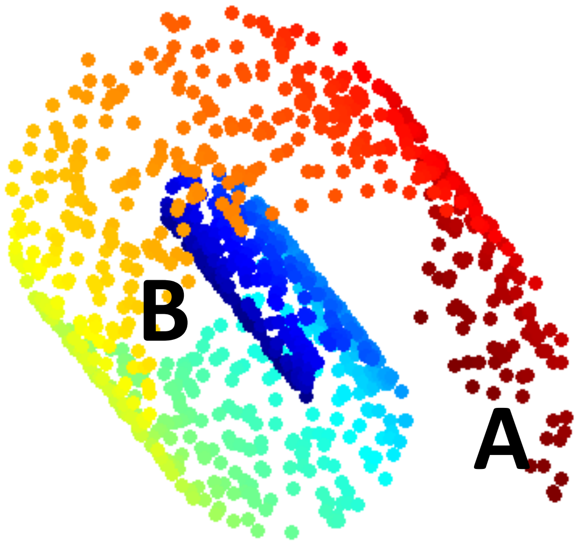
Manifold Learning

In which space should we put the "ball"?

- **Action description**
  - Can be very high-dimensional
  - Might be noisy
  - May lie on an intrinsically much lower dimensional space
- **Matching measure**
  - Can be adapted to the intrinsic structure of data
  - Can be learnt in a un/supervised manner

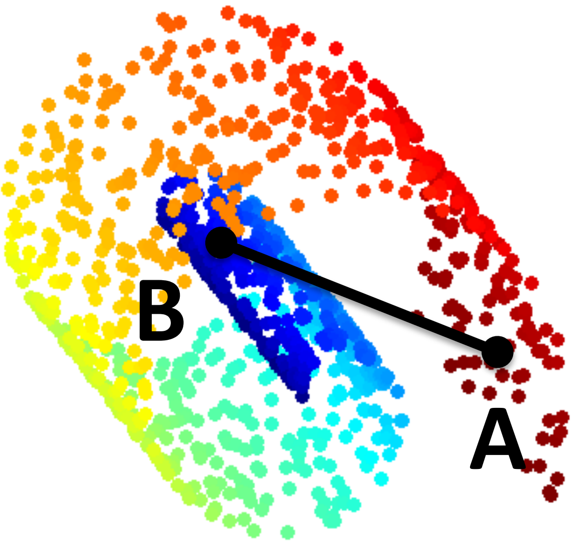
Manifold Learning

In which space should we put the "ball"?



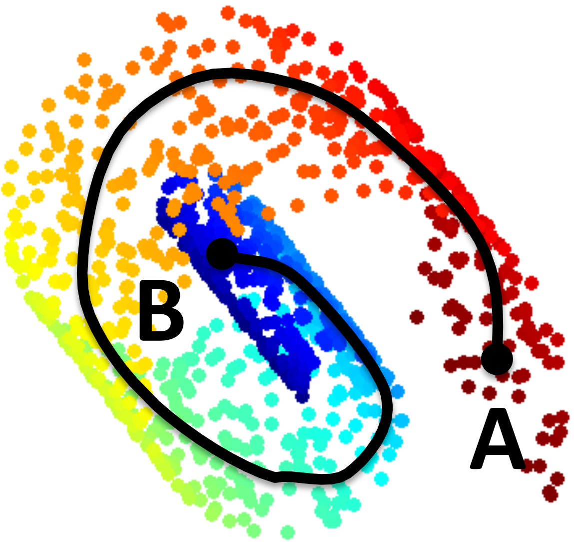
Manifold Learning

In which space
should we put the "ball"?



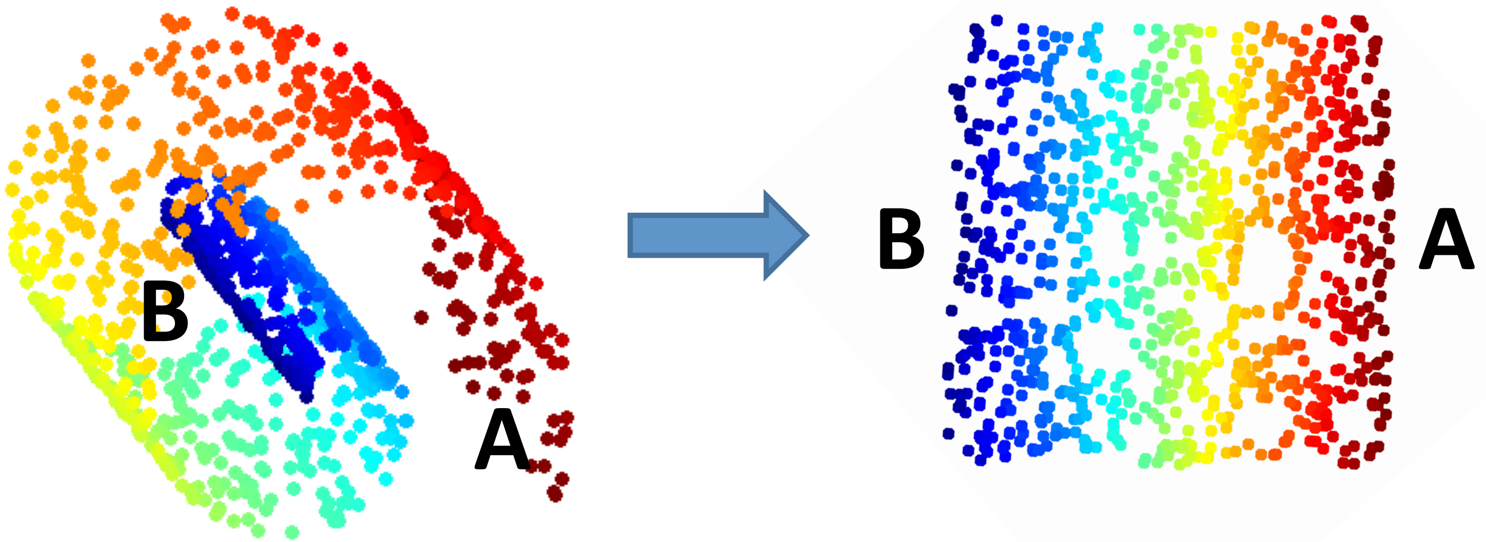
Manifold Learning

In which space should we put the "ball"?



Manifold Learning

In which space should we put the “ball”?



Manifold Learning

Apply the good old PCA

[Rosales 1998]
[Masoud and Papanikolopoulos 2003]

… or unravel a non-linear function between input and output spaces in an unsupervised way!

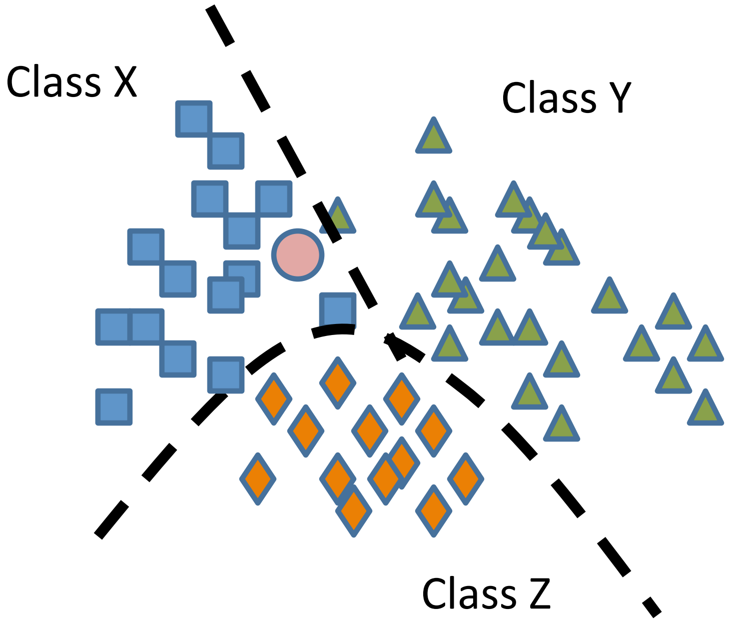
[Blackburn and Ribeiro 2007]
[Wang and Suter 2007]
[Wang and Suter 2008]

… or using some labeled data learn a metric between action instances discriminatively!

[Jia et al. 2008]
[Poppe and Poel 2008]
[Tran et al. 2008]

Discriminative Classifiers

[Smith et al. 2005]
[Jhuang et al. 2007]
[Laptev et al. 2007]
[Nowozin et al/ 2007]
[Fathi et al. 2008]



Given a pattern description, discriminative classifiers <u>focus on separating two or more classes,</u> rather than modeling the class-conditionals.

They constitute proxies to estimate the posterior probability.

Many off-the-shelf implementations available:
SVM, AdaBoost and variants, Random Forests

SVM
- Directly minimize a regularized upper bound on empirical classification error: Exact solution (QP)
- Generalizes well provided enough data
- Good with fixed vectorial description

Boosting
- Combine several weak classifiers into a strong one
- Ability to choose features
- Generalizes well provided enough data
- Blueprint algorithm: works with any weak learner/feature

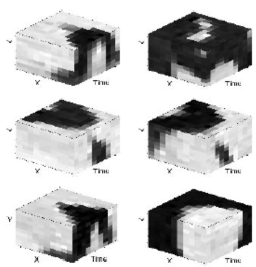
Random Forests
- Randomized extension of combined trees
- Ability to choose features
- Can seamlessly employ different types of features
- "A la mode"

Discriminative Classifiers

STIPs + BoW-based Action Recognition Framework

Detection of feature / interest points

Space-time patches

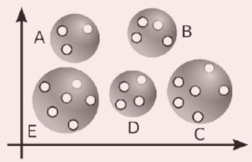


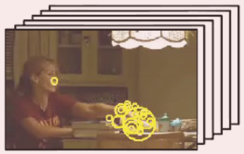
Description of space-time patches

Patch representation as feature vector $v = (v_1, v_2, \ldots, v_n)$

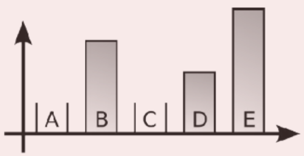
NN , SVM, Adaboost or RDF classifier

Training feature vectors are clustered with k-means (k=4000)





Each feature vector is assigned to its closest cluster center (visual word)



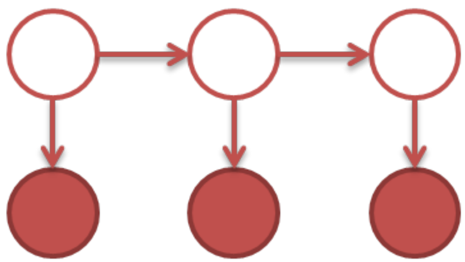
An entire video sequence is represented as occurrence histogram of visual words

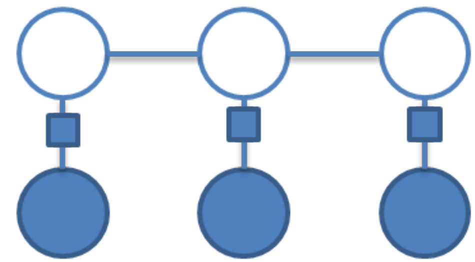
Feature Detectors
Harris3D, Hessian, Cuboid, … , Dense
Feature Descriptors
HOG/HOF, HOG3D, Cuboid, Ext. SURF, …

[Schuldt et al. 2004]
[Laptev 2005]
[Dollár et al. 2005]
[Oikonomopoulos et al. 2009]

State-Space Models

An action class and its observations can be described as a sequential probabilistic graphical model





Generative

P(C,D) then P(C|D) α P(D|C)P(C)

Hidden Markov Models (HMM)

generate states and observations

Discriminative

P(C|D) directly

Conditional Random Fields (CRF)

focus on the posterior without generating the states

Observations

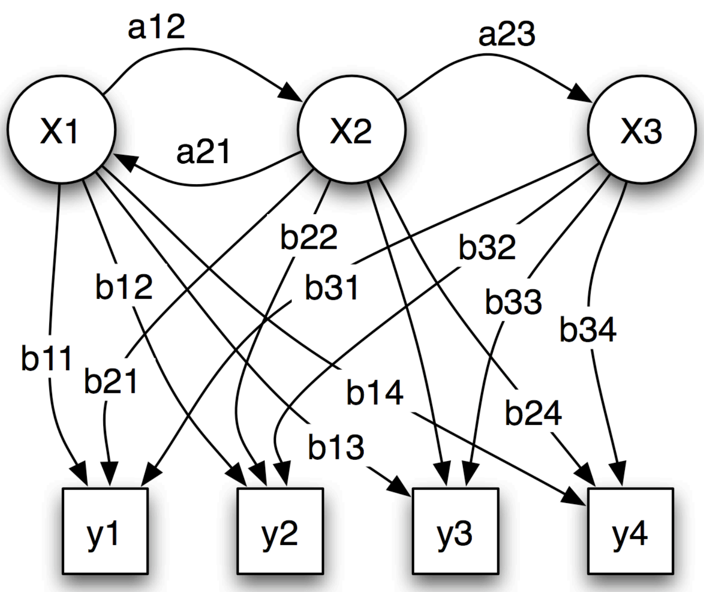
Sequence of visual descriptions of an action instance

States

Sequence of phases that an action instance undergoes

State-Space Models

Hidden Markov Models



X_i : hidden states
y_i : observations
a_ij : state transition probabilities
b_ij : output probabilities

[Feng and Perona 2002]
[Ikizler and Forsyth 2008]
[Lv and Nevatia 2006]
[Ramanan and Forsyth 2003]
[Yamato et al. 1992]

State-Space Models

Conditional Random Fields (CRF)

-Advantages over HMM-

- CRFs specify the probabilities of possible label sequences given an observation sequence:

  → **No modeling effort on the observations**

- The conditional probability of the label sequence can depend on arbitrary features of the observation sequence without requiring to account for the extra distributions:

  → **Can incorporate more information without extra effort**

  → **Independence assumptions not as strict as in HMMs**

[Ning et al. 2008]
[Shi et al. 2008]
[Sminchisescu et al. 2006]
[Wang and Suter 2007]

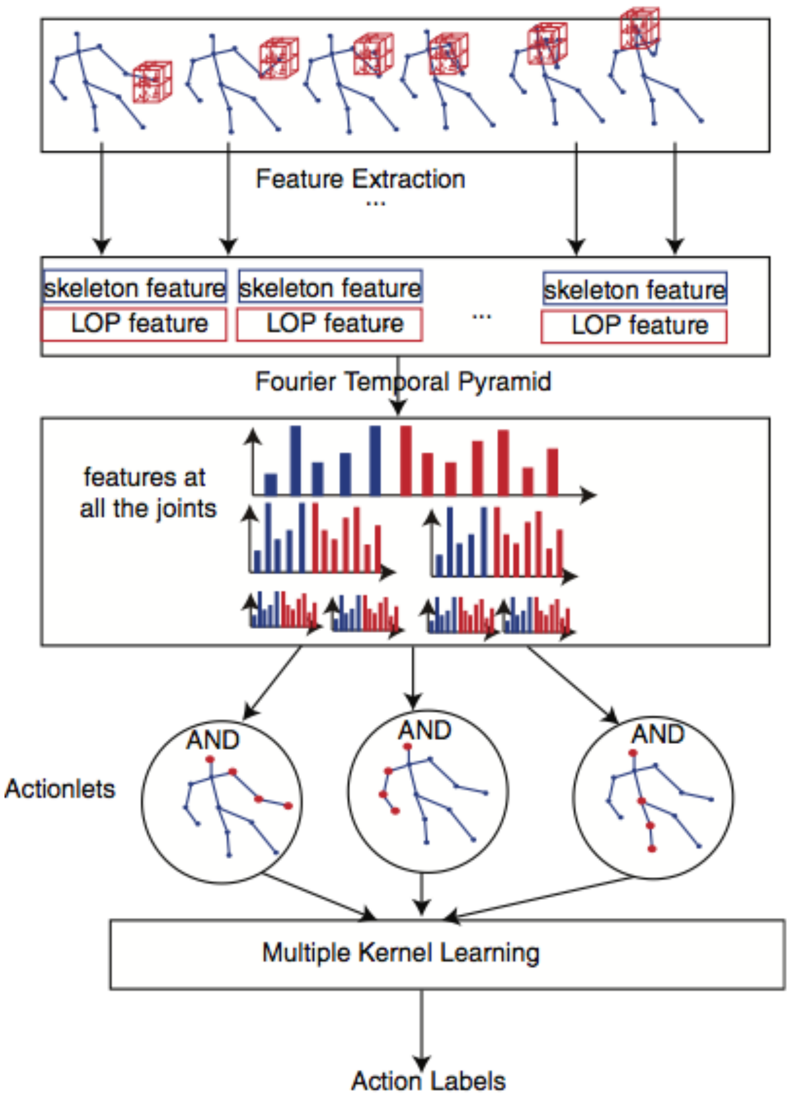
[Zhang and Gong 2010]
[Natarajan and Nevatia 2008]
[Mendoza and Blanca 2008]

there is more to the story…

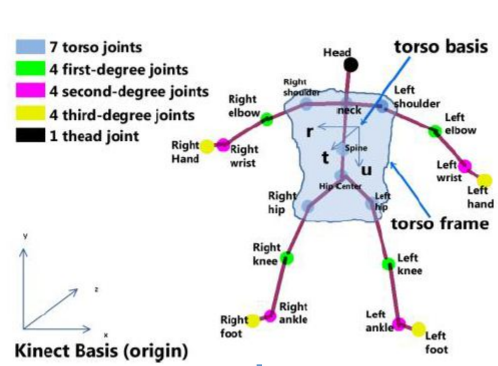
VARIATIONS ON THE THEME

Mining Action Data Using Context

Mining Action Data



[Wang et al. 2012]



266 feature time-series
13300 unique features

Discriminatively select features by RDF

SVM learning

Action classifiers
on selected features

[Negin et al. 2013]

Using Context – 1/4

Slide credit: Hedvig Kjellström

Using Context – 2/4

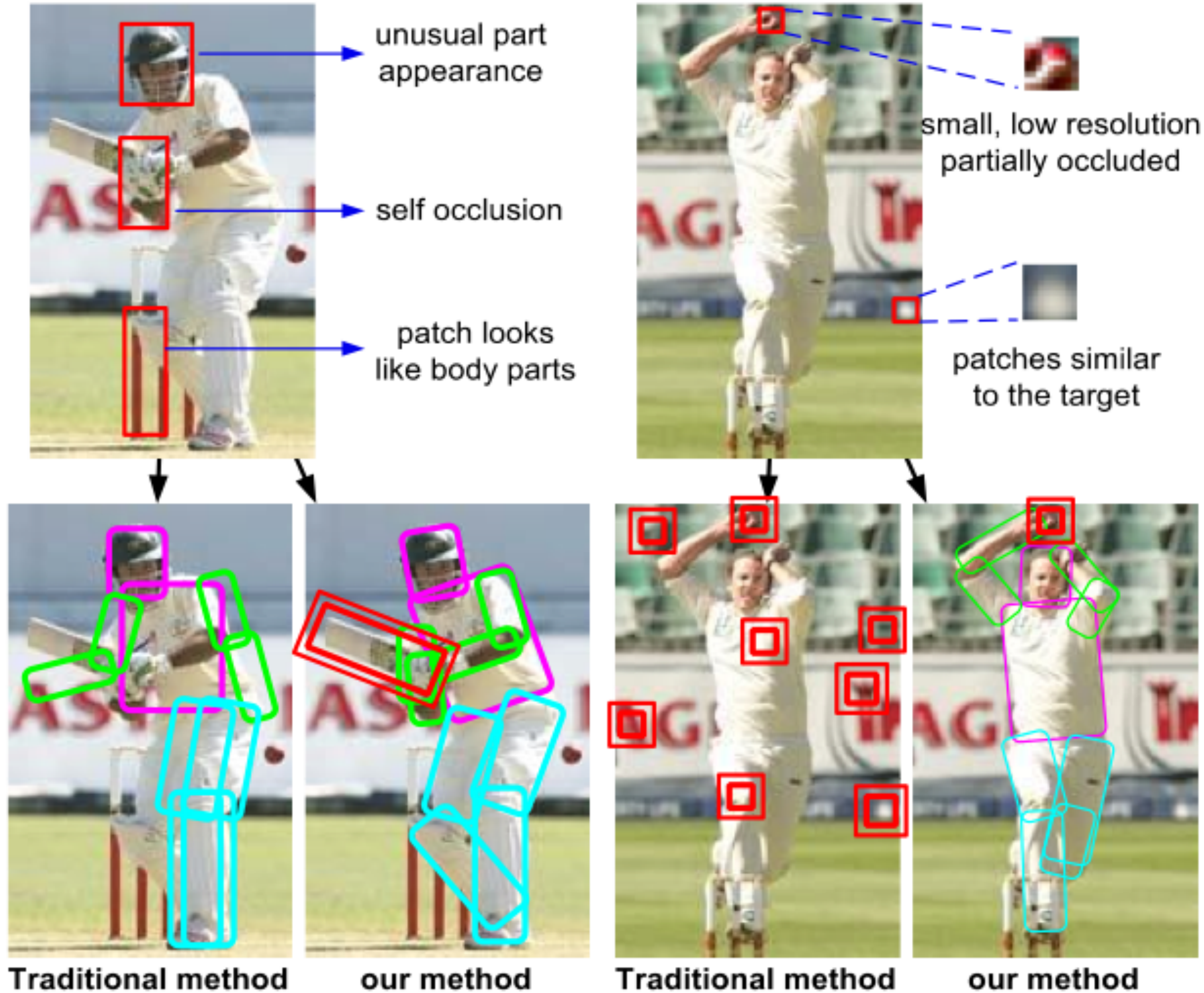
riding having-breakfast

- Object Context
  - The objects involved in the activity horse cup
  - Object state changes
- Scene Context
  - Scene category field kitchen
  - Scene topology, metrics
- Semantic Context
  - Grammars, temporally close actions activity structure
  - Speech, captions, storyline "Muybridge, race horse, 1887"
  - Expert and domain knowledge
- Photogrammetric Context
  - Image statistics, sensor info

Slide credit: Hedvig Kjellström

Using Context – 3/4

Object context
[Yao and Fei-Fei 2010]



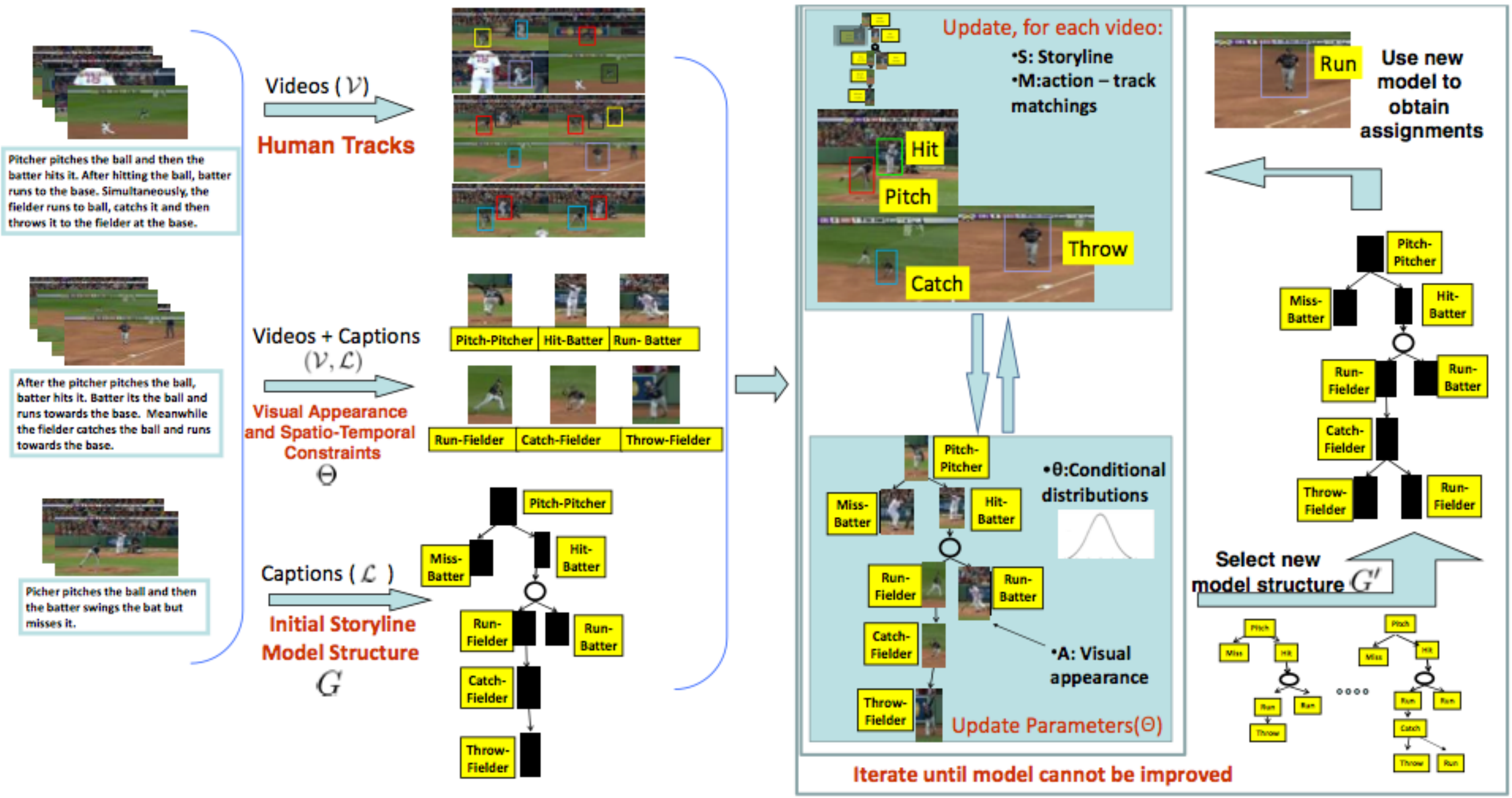
(a) Human pose estimation (b) Object (ball) detection

Slide credit: Hedvig Kjellström

Using Context – 4/4

Semantic context
[Gupta et al. 2009]

Slide credit: Hedvig Kjellström



what else?

CONCLUDING REMARKS

Challenges are still there…

CAN'T DO MUCH FOR THESE!

- ❑ Class Definitions and Variability
- ❑ Environment and Recording Settings
- ❑ Spatio-Temporal Variability

CAN AND SHOULD DO A LOT MORE HERE!

- ❑ Real-Time Recognition
- ❑ On-the-Fly Recognition
- ❑ Training Data Collection and Labeling
- ❑ Evaluation and Benchmarking

But the biggest (and most rewarding) ones are how to…

ADAPT DOMAINS
GO LARGE SCALE!

no need to thank ☺

REFERENCES

Dataset Links

The Usual Suspects

KTH: http://www.nada.kth.se/cvap/actions/
Weizmann: http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html
INRIA IXMAS: http://shivvitaladevuni.com/action_rec/ixmas_example.htm

The Wild Ones

Hollywood2 Dataset: http://www.di.ens.fr/~laptev/actions/hollywood2/
UCF Datasets: http://crcv.ucf.edu/data/UCF101.php
HMDB: http://serre-lab.clps.brown.edu/resources/HMDB/
ActionBank: http://www.cse.buffalo.edu/~jcorso/r/actionbank/

Surveillance Datasets

CAVIAR: http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/
Visor: http://imagelab.ing.unimore.it/visor/
PETS Datasets: http://www.hitech-projects.com/euprojects/cantata/datasets_cantata/dataset.html
SDHA 2010: http://cvrc.ece.utexas.edu/SDHA2010/

ADL Datasets

UCI ADL Dataset: http://deepthought.ics.uci.edu/ADLdataset/adl.html
TUM Kitchen Dataset: http://ias.cs.tum.edu/software/kitchen-activity-data
YouCook: http://www.cse.buffalo.edu/~jcorso/r/youcook/
UESTC Senior Home Monitoring: http://www.uestcrobot.net/senioractivity/

RGBD Datasets

Berkeley MHAD: http://tele-immersion.citris-uc.org/berkeley_mhad/
MSR Datasets: https://research.microsoft.com/en-us/um/people/zliu/actionrecorsrc/default.htm
http://research.microsoft.com/en-us/um/cambridge/projects/msrc12/
Cornell Activity Dataset: http://pr.cs.cornell.edu/humanactivities/data.php#format
LIRIS: http://liris.cnrs.fr/voir/activities-dataset/
WorkoutSU-10 Exercise: http://vpa2.sabanciuniv.edu/databases/WorkoutSU-10/

Survey References

Aggarwal, J. K., & Ryoo, M. S. (2011). Human activity analysis. *ACM Computing Surveys*, *43*(3), 1–43. doi:10.1145/1922649.1922653

Moeslund, T. B., Hilton, A., & Krüger, V. (2006). A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, *104*(2-3), 90–126. doi:10.1016/j.cviu.2006.08.002

Poppe, R. (2010). A survey on vision-based human action recognition. *Image and Vision Computing*, *28*(6), 976–990. doi:10.1016/j.imavis.2009.11.014

Turaga, P., Chellappa, R., Subrahmanian, V. S., & Udrea, O. (2008). Machine Recognition of Human Activities: A Survey. *IEEE Transactions on Circuits and Systems for Video Technology*, *18*(11), 1473–1488. doi:10.1109/TCSVT.2008.2005594

Weinland, D., Ronfard, R., & Boyer, E. (2011). A survey of vision-based methods for action representation, segmentation and recognition. *Computer Vision and Image Understanding*, *115*(2), 224–241. doi:10.1016/j.cviu.2010.10.002

Work References – 1/5

Nearest Neighbor Scheme

Dhruv Batra, Tsuhan Chen, Rahul Sukthankar, Space–time shapelets for action recognition, in: Proceedings of the Workshop on Motion and Video Computing (WMVC'08), Copper Mountain, CO, January 2008.

Lena Gorelick, Moshe Blank, Eli Shechtman, Michal Irani, Ronen Basri, Actions as space–time shapes, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 29 (12) (2007) 2247–2253.

Aaron F. Bobick, James W. Davis, The recognition of human movement using temporal templates, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 23 (3) (2001) 257–267.

Mikel D. Rodriguez, Javed Ahmed, Mubarak Shah, Action MACH: a spatio- temporal maximum average correlation height filter for action recognition, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08), Anchorage, AK, June 2008.

Zhe Lin, Zhuolin Jiang, Larry S. Davis, Recognizing actions by shape-motion prototype trees, in: Proceedings of the International Conference On Computer Vision (ICCV'09), Kyoto, Japan, September 2009.

Dynamic Time Warping

Ashok Veeraraghavan, Amit K. Roy-Chowdhury, Rama Chellappa, Matching shape sequences in video with applications in human movement analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 27 (12) (2005) 1896–1909.

Ashok Veeraraghavan, Rama Chellappa, Amit K. Roy-Chowdhury, The function space of an activity, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'06), vol. 1, New York, NY, June 2006, pp. 959–968.

Benjamin Yao, Song-Chun Zhu, Learning deformable action templates from cluttered videos, in: Proceedings of the International Conference On Computer Vision (ICCV'09), Kyoto, Japan, September 2009.

Work References – 2/5

Principal Component Analysis

Osama Masoud, Nikos Papanikolopoulos, A method for human action recognition, Image and Vision Computing 21 (8) (2003) 729–743.

Rómer E. Rosales, Recognition of human action using moment-based features, Technical Report BU-1998-020, Boston University, Computer Science, Boston, MA, November 1998.

Manifold Learning: Non-linear Methods

Jaron Blackburn, Eraldo Ribeiro, Human motion recognition using Isomap and dynamic time warping, in: Human Motion: Understanding, Modeling, Capture and Animation (HUMO'07), Lecture Notes in Computer Science, Rio de Janeiro, Brazil, October 2007, pp. 285–298 (Number 4814).

Liang Wang, David Suter, Learning and matching of dynamic shape manifolds for human action recognition, IEEE Transactions On Image Processing (TIP) 16 (6) (2007) 1646–1661.

Liang Wang, David Suter, Visual learning and recognition of sequential data manifolds with applications to human movement analysis, Computer Vision and Image Understanding (CVIU) 110 (2) (2008) 153–172.

Metric Learning

Kui Jia, Dit-Yan Yeung, Human action recognition using local spatio-temporal discriminant embedding, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08), Anchorage, AK, June 2008.

Ronald Poppe, Mannes Poel, Discriminative human action recognition using pairwise CSP classifiers, in: Proceedings of the International Conference on Automatic Face and Gesture Recognition (FGR'08), September 2008, Amsterdam, The Netherlands, 2008.

Du Tran, Alexander Sorokin, David A. Forsyth, Human activity recognition with metric learning, in: Proceedings of the European Conference on Computer Vision (ECCV'08) – part 1, Lecture Notes in Computer Science, Marseille, France, October 2008, pp. 548–561 (Number 5302).

Work References – 3/5

Discriminative Classifier Methods

Alireza Fathi, Greg Mori, Action recognition by learning mid-level motion features, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08), Anchorage, AK, June 2008.

Hueihan Jhuang, Thomas Serre, Lior Wolf, Tomaso Poggio, A biologically inspired system for action recognition, in: Proceedings of the International Conference On Computer Vision (ICCV'07), Rio de Janeiro, Brazil, October 2007.

Ivan Laptev, Patrick Pérez, Retrieving actions in movies, in: Proceedings of the International Conference On Computer Vision (ICCV'07), Rio de Janeiro, Brazil, October 2007.

Paul Smith, Niels da Vitoria Lobo, Mubarak Shah, TemporalBoost for event recognition, in: Proceedings of the International Conference On Computer Vision (ICCV'05), vol. 1, Beijing, China, October 2005, pp. 733–740.

Sebastian Nowozin, Gökhan Bakır, Koji Tsuda, Discriminative subsequence mining for action classification, in: Proceedings of the International Conference On Computer Vision (ICCV'07), Rio de Janeiro, Brazil, October 2007.

Space-Time Interest Points Based Framework

I. Laptev, On Space-Time Interest Points, ; in International Journal of Computer Vision, vol 64, number 2/3, pp.107-123, 2005.

I. Laptev, B. Caputo, C. Schuldt and T. Lindeberg, Local Velocity-Adapted Motion Events for Spatio-Temporal Recognition, in Computer Vision and Image Understanding, 108:207-229, 2007.

Christian Schuldt, Ivan Laptev and Barbara Caputo, Recognizing Human Actions: A Local SVM Approach in Proc. ICPR'04, Cambridge, UK, pp.III:32--36, 2004.

Piotr Dollár, Vincent Rabaud, Garrison Cottrell and Serge Belongie, Behavior Recognition via Sparse Spatio-Temporal Features, in ICCV VS-PETS 2005, Beijing, China.

Antonios Oikonomopoulos, Maja Pantic, Ioannis Patras, Sparse B-spline polynomial descriptors for human activity recognition, Image and Vision Computing 27(12) (2009) 1814–1825.

Work References – 4/5

State-Space Methods: Generative

Xiaolin Feng, Pietro Perona, Human action recognition by sequence of movelet codewords, in: Proceedings of the International Symposium on 3D Data Processing, Visualization, and Transmission (3DPVT'02), Padova, Italy, June 2002, pp. 717–721.

Nazlı Ikizler,DavidA.Forsyth,Searchingforcomplexhumanactivitieswith no visual examples, International Journal of Computer Vision (IJCV) 30 (3) (2008) 337–357.

Fengjun Lv, Ram Nevatia, Recognition and segmentation of 3-D human action using HMM and multi-class adaBoost, in: Proceedings of the European Conference on Computer Vision (ECCV'06), Lecture Notes in Computer Science, vol. 4, Graz, Austria, May 2006, pp. 359–372 (Number 3953).

Deva Ramanan, David A. Forsyth, Automatic annotation of everyday movements, in: Advances in Neural Information Processing Systems (NIPS), vol. 16, Vancouver, Canada, 2003.

Junji Yamato, Jun Ohya, Kenichiro Ishii, Recognizing human action in time- sequential images using hidden Markov model, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'92), Champaign, IL, June 1992, pp. 379–385.

State-Space Methods: Discriminative

Huazhong Ning, Wei Xu, Yihong Gong, Thomas S. Huang, Latent pose estimator for continuous action recognition, in: Proceedings of the European Conference on Computer Vision (ECCV'08) – part 2, Lecture Notes in Computer Science, Marseille, France, October 2008, pp. 419–433 (Number 5305).

Qinfeng Shi, Li Wang, Li Cheng, Alex Smola, Discriminative human action segmentation and recognition using semi-Markov model, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08), Anchorage, AK, June 2008.

Cristian Sminchisescu, Atul Kanaujia, Dimitris N. Metaxas, Conditional models for contextual human motion recognition, Computer Vision and Image Understanding (CVIU) 104 (2–3) (2006) 210–220.

Liang Wang, David Suter, Recognizing human activities from silhouettes: motion subspace and factorial discriminative graphical model, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'07), Minneapolis, MN, June 2007.

Work References – 5/5

State-Space Methods: Discriminative – *cont'd*

Jianguo Zhang, Shaogang Gong, Action categorization with modified hidden conditional random field, Pattern Recognition 43 (1) (2010) 197–203.

Pradeep Natarajan, Ram Nevatia, View and scale invariant action recognition using multiview shape-flow models, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08), Anchorage, AK, June 2008.

M. Ángeles Mendoza, Nicolás Pérez de la Blanca, Applying space state models in human action recognition: a comparative study, in: International Workshop on Articulated Motion and Deformable Objects (AMDO'08), Lecture Notes in Computer Science, Port d'Andratx, Spain, July 2008, pp. 53–62 (Number 5098).

Mining Action Data

Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan, Mining actionlet ensemble for action recognition with depth cameras, CVPR 2012, page 1290-1297, Providence, USA, 2012.

F. Negin, F. Özdemir, C. B. Akgül, K. A. Yüksel, A. Erçil, A Decision Forest Based Feature Selection Framework for Action Recognition from RGB-Depth Cameras. Special Session on Recent Advances on RGB-D Camera Applications, International Conference on Image Analysis and Recognition (ICIAR 2013).

Using Context

Yao, Fei-Fei, "Modeling mutual context of object and human pose in human- object interaction activities", IEEE Conference on Computer Vision and Pattern Recognition 2010

Gupta, Srinivasan, Shi, Davis, "Understanding videos, constructing plots: Learning a visually grounded storyline model from annotated videos", IEEE Conference on Computer Vision and Pattern Recognition 2009.